TreeGenes: Past, Present, and Future Development

Jill Wegrzyn
Plant and Animal Genome Conference
January 14th 2019
Dendrome/TreeGenes was the original three USDA-ARS databases
So much time...so many technologies!

So much time...so many data types!

1991: Dendrome (stand-alone software)
1995: Database Online (ACEDB)
2001: Oracle Database
2005: MySQL
2009: PostGreSQL
2015: Tripal Framework

Colleague Directory Literature Database
Markers (SSR, AFLP) Genetic Maps, Community Markers (SSR, AFLP, RAPD) ESTs Genetic Maps, Community Markers, Community ESTs cDNAs Genetic Maps, Markers, Community ESTs cDNAs Phenotypes Genetic Maps, Markers, Community ESTs cDNAs Phenotypes SNPs Genomes

dendrome.ucdavis.edu

Treegenes.ucdavis.edu
Treegenesdb.org
TreeGenes Audience

- Academic Research Labs
- Non-profit research institutions
- US Forest Service
- Industry
  - Timber companies
- JGI/DOE
  - Data exchange
- Genbank
  - Data exchange
TreeGenes Staffing

![Graph showing staffing trends over years]

- **FTE**
- **Undergrads**
- **Consultants**

Years (1991-2018)

Staffing levels fluctuate over the years, with a peak around 2008 and a decline in 2014.
BT Era (Before Tripal)
# Primary Structure

## Web
- Custom Frontends
- Search Fields/Summary Pages
- Advanced Interface (DiversiTee)

## Database
- Custom Schema
- GMOD components:
  - Genetic Maps (CMAP)
  - Genomes (Gbrowse/Jbrowse)

## Scope
- Genetic Maps
- Sequence
- Phenotypes
- Genotypes
- Conifers/Forest Trees

## Funding
- No *dedicated* funding
  - Forest Service
  - USDA CAP
  - NSF Projects
  - Inbetween? plugging along (i.e. Maintenance)

---

**Welcome to the Generic Model Organism Database** project, a collection of open source software tools for managing, visualising, storing, and disseminating genetic and genomic data.

### Get Started
Read the GMOD overview for the big picture, or visit GMOD Components for a comprehensive list of GMOD tools. If GMOD looks promising for your needs, consider attending the next GMOD community meeting.

### Get Support
GMOD support is available from several different sources. Support introduces each support option (this web site, GMOD Mailing Lists, Training and Outreach activities (including GMOD Schools), and the GMOD Help Desk) and offers guidance on which one is the most appropriate for your question.

### Get Involved
As an open source project GMOD relies on the donation of time and software by groups and individuals. Contribution of new tools, adoption of existing ones, and improving the documentation are all welcome. Existing and potential users are encouraged to provide feedback via mailing lists or the help desk. You can also attend project meetings.
AT Era (After Tripal)
Current Funding

• NSF PGRP (2015-2019)
  • Conversion to Tripal
  • Ontologies
  • CartograTree
• NSF DIBBS (2014-2018)
  • Tripal Galaxy
  • Module development
  • CartograTree
• USDA AFRI - PineSNP
  • Partial project support
Tripal v3 Framework in TreeGenes

- Plant Computational Genomics Lab
- Department of Ecology & Evolutionary Biology
- UConn

Tripal v3 Framework in TreeGenes

- **Web**: JBrowse, CartograTree, OrthoQuery, TSeq
- **Database**: CHADO + Web Services
- **Analysis**: Galaxy

Copyright 2019-2023 UConn

Plant Computational Genomics Lab | Department of Ecology & Evolutionary Biology | UConn
1,850 species from 112 genera
  – At least one genetic artifact from each species
Full genome sequence: 35 species
Transcriptome/Expression resources: 6,920,817 sequences from 322 species
101 genetic maps from 37 species
Population studies
  – Georeferenced trees
  – Extensive genotypic (GBS and array) and phenotypic data
Relational Data in TreeGenes

- Literature
- Sequences:
  - Genome
  - Scaffolds
  - Annotations
  - Transcriptome
  - TSA
  - EST
  - cDNA
  - Unigene
  - Marker
  - Protein
  - SNPs

- Colleagues
- Environmental Variables and Range maps
- Phenotype
- Genotype

1KP

NCBI
TreeGenes Database: Users

3,100 unique visitors from 116 countries

Unique Web Visitors to TreeGenes (by month), July-December 2018
Community Resources: Colleague directory, mailing list, meetings, employment listings, and literature
Welcome to TreeGenes
You're now logged in and ready to go!

- Add/Edit Profile
- Add/Edit Workshop/Conference
- Add/Edit Employment Listings

Free account to: edit profile, add listings, submit data!
# Eucalyptus grandis

## Summary

<table>
<thead>
<tr>
<th><strong>Species:</strong></th>
<th>Eucalyptus grandis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Common Name:</strong></td>
<td>Grand Eucalyptus</td>
</tr>
<tr>
<td><strong>TreeGenes Code:</strong></td>
<td>Eugr</td>
</tr>
<tr>
<td><strong>Lineages:</strong></td>
<td>Viridiplantae; Streptophyta; Streptophytina; Embryophyta; Tracheophyta; Euphyllophyta; Spermatophyta; Magnoliophyta; Mesangiospermae; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Myrtales; Myrtaceae; Myrtoideae; Eucalyptaeae; Eucalyptus</td>
</tr>
<tr>
<td><strong>Haploid Chromosome Number:</strong></td>
<td>12</td>
</tr>
<tr>
<td><strong>GBIF:</strong></td>
<td><a href="https://www.gbif.org/species/3177028">https://www.gbif.org/species/3177028</a></td>
</tr>
<tr>
<td><strong>Bioproject Records:</strong></td>
<td>50</td>
</tr>
<tr>
<td><strong>Biosample Records:</strong></td>
<td>130</td>
</tr>
</tbody>
</table>
TreeGenes: Genome Access

- **Betula nana**
  - Version: v1.0
  - Organism Detail | JBrowse
  - Genome FASTA
  - Annotation (GFF3)

- **Carica papaya**
  - Version: v0.4
  - Organism Detail | JBrowse
  - Genome FASTA
  - Annotation (GFF3)
  - Genes (FASTA)
  - Proteins (FASTA)

- **Elaeis guineensis**
  - Version: v1.0
  - Organism Detail | JBrowse
  - Genome FASTA
  - Annotation (GFF3)
  - Genes (FASTA)
  - Proteins (FASTA)

- **Eucalyptus camaldulensis**
  - Version: v1.0
  - Organism Detail | JBrowse
  - Genome FASTA
  - Annotation (GFF3)
  - Genes (FASTA)
  - Proteins (FASTA)

- **Eucalyptus grandis**
  - Version: v1.0
  - Organism Detail | JBrowse
  - Annotation (GFF3)
  - Genes (FASTA)
  - Proteins (FASTA)

- **Fraxinus excelsior**
  - Version: v0.5
  - Organism Detail | JBrowse
  - Genome FASTA
  - Annotation (GFF3)
  - Genes (FASTA)
  - Proteins (FASTA)

- **Ginkgo biloba**
  - Version: v1.0
  - Organism Detail | JBrowse
  - Genome FASTA
  - Annotation (GFF3)
  - Genes (FASTA)
  - Proteins (FASTA)

- **Gnetum montanum**
  - Version: v1.0
  - Organism Detail | JBrowse
  - Genome FASTA
  - Annotation (GFF3)
  - Genes (FASTA)
  - Proteins (FASTA)
Tripal v3 Framework in TreeGenes

- Plant PopGen Submit (TPPS)
- JBrowse
- CartograTree
- OrthoQuery
- TSeq

Web
- CHADO + Web Services

Analysis
- Galaxy
TSeq

Diamond search
BLASTp/x
Faster performance
BLAST
BLASTn
Analysis on remote (non-web) server
TreeGenes UniGenes to OrthoQuery

### TreeGenes: OrthoQuery Module

#### 3 Primary Uses Cases for the OrthoQuery Module

**Input: Small Set of Protein Sequences**
- **Question:** Which gene family does it belong to?
- **Process:**
  1. Map existing orthogroups to the protein sequence(s)
  2. Provide visualization
  3. Assign functional annotation

**Input: Transcriptome/Gene Set**
- **Question:** Orthologous genes, expansion/contraction of gene families
- **Process:**
  1. Construct Proteome
  2. Discover new orthogroups with existing tools
  3. Provide visualization & statistical results

**Input: Proteome**
- **Question:** Orthologous genes, expansion/contraction of gene families
- **Process:**
  1. Discover new orthogroups with existing tools
  2. Provide visualization & present statistical results

[DEMO] [CONTINUE TO ORTHOQUERY]
Tripal Plant PopGen Submit (TPPS)

Population Study
- Publication
- Species

Study Design
- Landscape
- Common Garden
- Greenhouse
- Growth Chamber
- Breeding (Plot)

Phenotype, Genotype, Environment
- Georeferenced

Raw Data
Minimal Information About a Plant Phenotyping Experiment (MIAPPE)

MIAPPE

Minimum Information About a Plant Phenotyping Experiment

MIAPPE is an open, community driven project to harmonize data from plant phenotyping experiments. MIAPPE comprises both a conceptual checklist of metadata required to adequately describe a plant phenotyping experiment, as well as a suite of tools to validate, store and disseminate MIAPPE-compliant data. We welcome contributions from all those interested in plant phenotyping.

Development of MIAPPE is an open process, so if you would like to do more than just comment, and to participate in the ongoing development, please let us know via miappe-feedback@ebi.ac.uk.

MIAPPE v1.1

MIAPPE version 1.1 was officially released on 10th January 2019, following consideration of responses to two requests for comments. Major developments over v1.0 include:

- Integration of ontologies:
  - Plant Ontology (PO)
  - Crop Ontology (CO)
  - Trait Ontology (TO)
  - Environmental Ontology (ENVO)
  - Chemical Ontology (CHEBI)
TreeGenes Ontology Integration

TreeGenes Phenotype and Structure Ontology Directory

TREEGENES PHENOTYPE AND STRUCTURE ONTOLOGY DIRECTORY

TreeGenes is actively curating submitted traits and structures and assigning these traits to the following ontologies: Plant Ontology (structure), Trait Ontology (traits), Crop Ontology (woody trait ontology), and PATO (descriptive ontology). When possible, assignments are made to existing terms that are curated by the PlantGnome and Crop Ontology projects. For terms that are not yet assigned to Assignment of individual measurements and structures to an ontological framework provides improved recovery of data downstream for all users. This also enables comprehensive analytical pipelines that can efficiently integrate across a multitude of independent studies. To submit your phenotype and/or genotype study, please use the [Submit Pipeline](#).

We currently have **306** unique phenotypes described on **22** unique structures representing a total of **724,265** phenotypic measures from **414,511** samples representing **18** species.

ONTOGONY TERM SEARCH

![Plant Ontology (PO)](image1)

![Trait Ontology (TO)](image2)

![Crop Ontology (CO)](image3)
**Tripal Plant PopGen Submit (TPPS)**

Phenotype Metadata File: Please upload a file containing columns with the name, attribute, description, and units of each of your phenotypes.

- Phenotype Metadata.xlsx
- File Upload empty field: NA

By default, TPPS will treat cells with the value "NA" as empty. If you used a different empty value indicator, please provide it here.

**DEFINE DATA**

Please define which columns hold the required data: Phenotype name

<table>
<thead>
<tr>
<th>Phenotype Name/Identifier</th>
<th>Attribute</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>phenotype 1</td>
<td>age</td>
<td>years</td>
</tr>
<tr>
<td>phenotype 2</td>
<td>age</td>
<td>years</td>
</tr>
<tr>
<td>phenotype 3</td>
<td>age</td>
<td>years</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Phenotype file below.**

- Phenotype 2 Attribute: ¨
- Some examples of attributes include: 'amount', 'width', 'mass density', 'area', 'height', 'age', 'broken', 'time', 'color', 'composition', etc.
- Phenotype 2 Description: 
- Please provide a short description of Phenotype 2
- Phenotype 2 Units: 

TreeGenes Data Repository Accession (TGDR#####) -> DOI
# TreeGenes to CartograTree

## TPPS/TGDR DETAILS FOR TGDR001

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession</td>
<td>TGDR001</td>
</tr>
<tr>
<td>Title</td>
<td>Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem.</td>
</tr>
<tr>
<td>Species</td>
<td>Populus trichocarpa</td>
</tr>
<tr>
<td>Study Type</td>
<td>GxP</td>
</tr>
<tr>
<td>File Downloads</td>
<td>ASSOCIATION RESULTS FILE 1 FILE</td>
</tr>
<tr>
<td></td>
<td>GENOTYPES SNP</td>
</tr>
<tr>
<td></td>
<td>GPS COORDINATES</td>
</tr>
<tr>
<td></td>
<td>HAPLOTYPE DATA FILE</td>
</tr>
<tr>
<td></td>
<td>PHENOTYPES</td>
</tr>
<tr>
<td></td>
<td>PHENOTYPES DEFINITIONS</td>
</tr>
</tbody>
</table>

**CartograTree**

- [View in CartograTree](#)

<table>
<thead>
<tr>
<th>Tree Count</th>
<th>1376</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype Count</td>
<td>4032</td>
</tr>
<tr>
<td>Unique Phenotypes</td>
<td>3</td>
</tr>
</tbody>
</table>
CartograTree: Integrating environmental layers with georeferenced trees
CartograTree: Integrating environmental layers with georeferenced trees
CartograTree: Save searches locally and select for meta-analysis
Bringing Analytical Capacity to the Data

CartograTree

Tree Genes

Hardwood Genomics

Genome Database for Rosaceae
Galaxy: Open Source Web-based Platform for Bioinformatic Analysis
Workflows for Landscape Genomics: Integrating across diverse datasets

Genotypic data

Galaxy

Environmental data

Correlated markers
Workflows: Executed in Galaxy with metadata

Pre-processing

- Select G x E data
- Filter by individuals
- Filter by SNPs
- Filter by MAF
- Filter by MGF
- Reduce dimensionality

Preliminary analysis

- Landscape genetics analysis
- Population structure
- Heterozygosity

Display results
Sustainability?
Sustainability

Funding
- CartograTree
- Modules + APIs
- Tripal Core

Collaborative
- Consultant
- DMP Coordination

Industry/NGO/Gov
- Industry partners?
- GTTN
- Forest Service
Global Timber Tracking Network

FIGHT AGAINST ILLEGAL LOGGING
Meet our network