Beyond Search and Display – Analyze Tripal Data with CartograTree

Sean Buehler
Plant and Animal Genome Conference
January 13, 2019
Tripal Possibilities

• Tripal is great for biological data!
  • Storage
  • Searching
  • Displaying

• Community Modules
• Tripal 3
• Forest Trees - Nearly 1800
• Very few with reference genome - 40
• Data from Population Studies
  • Phenotypic values
  • Genotypic data
  • Environmental data
• Many georeferenced trees
Going beyond search and display

What can we do with all this data?
Analysis!
What kinds?

• Association Genetics (Phenotypic and Genotypic)
  • Genotype contribution to traits (timber production, pests & pathogens)

• Landscape Genetics (Genotypic and Environmental)
  • Genotypes are most adapted to specific elevations & climates
  • Individual suitability for assisted migration
The current process

<table>
<thead>
<tr>
<th>Process</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Identify public sources of genotypic and phenotypic</td>
<td>Various sites in different formats</td>
</tr>
<tr>
<td>2) Get geo-referenced environmental values for that data</td>
<td>Tedious</td>
</tr>
<tr>
<td>3) Installing software and managing packages on HPC</td>
<td>Many dependencies, constantly changing versions</td>
</tr>
<tr>
<td>4) Upload data to HPC</td>
<td>Large amount of data over network</td>
</tr>
<tr>
<td>5) Run analysis</td>
<td>Interacting with HPC</td>
</tr>
<tr>
<td>6) Analyze results</td>
<td>Non-trivial task</td>
</tr>
</tbody>
</table>
Solution?

CartograTre + TPPS
CartograTree

- Web-based map-driven
  - Genotypic
  - Phenotypic
  - Environmental
How it Works

• CartograTree is a Tripal extension module
  • PHP
  • NodeJS - API
  • MapBox
  • Geoserver

• Access existing data from the Tripal site (CHADO)
• Access environmental data from public repositories
• Integrate data and perform analysis through Galaxy
General Overview

G & P Data Sources
- TPPS
- DRYAD
- TreeSnap
- Other

Database
- TreeGenes

Computation and Heavy Lifting
- CartograTree

Environmental Layers
- Try-DB
- WorldClim
- US Forest Service
- CGIAR-CSI
- Additional Sources

Additional Sources
- Environmental Layers
- Database
- Computation and Heavy Lifting
Clade/Model Organism Database

- G & P Data Sources
  - TPPS
  - DRYAD
  - TreeSnap
  - Other

- Database
  - TreeGenes

- Computation and Heavy Lifting
  - CartograTree

- Environmental Layers
  - Try-DB
  - WorldClim
  - US Forest Service
  - CGIAR-CSI
  - Additional Sources

- Galaxy
Tripal Plant PopGen Submit Pipeline

What it is
- Short yet comprehensive series of forms
- Form adapts to the user’s input

Motivation
- Metadata surrounding a study:
  - Inconsistent
  - Sometimes lost after publication
Tripal Plant PopGen Submit

Population Study
- Publication
- Species

Study Design
- Landscape
- Common Garden
- Breeding (plot)

Tree Locations
- Geolocation

Raw Data
- Phenotype
- Genotype

A focus on metadata, georeferenced data
TreeGenes: lack of reference genome?
No problem!
TPPS in Action

[Image of the TPPS in Action page, showing a form with fields for species, treatments, and phenotypes, along with a map of gene expression data.]
Minimum Information About a Plant Phenotyping Experiment

Ontologies make everything succinct
- Plant
- Crop
- Trait
Dryad and TreeSnap

• Dryad - Repository of DOIs
  • Extract data from flat files

• TreeSnap
  • Data collected by citizen scientists
What the data looks like
Environmental Layers

G & P Data Sources
- TPPS
- DRYAD
- TreeSnap
- Other

Database
- TreeGenes

Computation and Heavy Lifting

Environmental Layers
- Try-DB
- WorldClim
- US Forest Service
- CGIAR-CSI
- Additional Sources
Environmental Layers

- WorldClim
  - Precipitation
  - Temperature
  - Solar radiation
  - Wind speed
  - Water Vapor
- Conservation Biology Institute
  - Major Soil Groups
- US Forest Service
  - Species range maps
- CGIARCSI
- Aridity
- Potential Evapotranspiration (PET)
- Solar radiation
- NEON
  - Remote sensing
- Other
  - Tree/Land cover
  - Canopy height
  - Forest zone
Environmental Layers
Environmental Layers

WorldClim – Global Climate Data
Environmental Layers

- WorldClim v2
  - Precipitation (WorldClim v2)
  - Temperature (WorldClim v2)
  - Solar Radiation (WorldClim v2)
  - Wind Speed (WorldClim v2)
  - Water Vapor (WorldClim v2)
    - Water vapor pressure
      - January
      - Opacity 74%

- Major Soil Types (Conservation Biology Institute)

- Major Soil Groups
  - Major Soil Groups
    - Opacity 58%

- Species Range Maps

- Land Cover
Environmental Layers
Species Range
CartograTree - Searches

Searches saved to User Profile

- Rerunning past searches
  - Different parameters
  - Different analysis
- Combining past searches
Analysis Stage

G & P Data Sources
- TPPS
- DRYAD
- TreeSnap
- Other

Database
- TreeGenes

Computation and Heavy Lifting
- CartograTree

Environmental Layers
- Try-DB
- WorldClim
- US Forest Service
- CGIAR-CSI
- Additional Sources
Preparing data for analysis

1. Filter chosen trees/data
   - Missing data

2. Choose appropriate workflow from Galaxy
Example workflow: Landscape Genomics
Conclusion

- Building on the functionality of Tripal
  - Phenotypic, genotypic, environmental + metadata
  - community modules
- TPPS - A pipeline to collect standardized and necessary data
- CartograTree - A web-based tool to integrate data and launch analyses
Future Work

• More Galaxy workflows
• NEON data as environmental layers
• Portability to other Tripal sites
  • Pair TPPS and CartograTree in deployment
• Display analysis results as a layer on map
University of Connecticut
Nic Herndon
Risharde Ramnath
Ronald Santos
Peter Richter
Taylor Falk
Emily Grau
Jill Wegrzyn

Washington State University
Stephen Ficklin
Dorrie Main
Sook Jung

University of Tennessee
Margaret Staton

Poster: PO0527

@TreeGenes