Galaxy Enables Integrated Analysis of Phenotypic, Genotypic, and Environmental Data for Geo-referenced Trees in CartograTree

Nic Herndon, Peter Richter, Taylor Falk, Sean Buehler, Emily Grau, Dorrie Main, Sook Jung, Stephen Ficklin, Margaret Staton, Alex Feltus, and Jill Wegrzyn

GCC BOSC 2018
Motivation
Motivation

Genotype

Environment

Phenotype
CartograTree’s user interface

Other searchable fields: publication, study type, phenotypes, genotype (marker type), sequence source
CartograTree’s main data source: TreeGenes DB
TreeGenes data sources

- NCBI
- Drupal
- TPPS
- PostgreSQL
- CHADO
- Tripal

Poster G08
Environmental data sources

- WorldClim - Global
  Free climate data for ecological models

- Harmonized World Data Base
  Uploaded by Conservation Biology

- Little's Range and
  RWU NE-4153, USDA Forest Service, Northeastern PA

- CartograTree
  A Forest Tree Map Utility

- Species Table
  TARIS, USDA Forest Service, Northeastern Research Station, Delaware, Ohio, USA

- CartograTree
  A Forest Tree Map Utility
  Welcome to CartograTree - Log in to get started!
  23,860 trees from 36 species.
Environmental data sources

WorldClim - Global Climate Data
Free climate data for ecological modeling and GIS

Harmonized World Soil Dataset - Major Soil Groups
Uploaded by Conservation Biology Institute

Little's Range and FIA Importance Value Distribution Maps
Spatial database for 135 eastern US tree species
Anantha M. Prasad and Louis R. Iverson

Node.js
OpenLayers
PostgreSQL
CHADO
GeoServer
Analysis with Sambada

Genotypic data

Sambada

Correlations

Environmental data
1. Select G+E data with CartograTree
1. Select G+E data with CartograTree

- Data compatible for (meta-)analysis
  - Same marker types (e.g., SNPs v. SSRs)
  - Evidence for SNPs – quality score, coverage
  - Same –ome (e.g., genome v. transcriptome)
  - Same assembly version
2. Filter genetic data

- Keep users in TreeGenes
- Expose only some parameters
2. Filter genetic data
2. Filter genetic data

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
<th>SNP4</th>
<th>SNP5</th>
<th>SNP6</th>
<th>SNP7</th>
<th>SNP8</th>
<th>SNP9</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREE1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TREE4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE5</td>
<td>1</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TREE7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TREE8</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Filter genetic data: missing SNPs
2. Filter genetic data

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
<th>SNP4</th>
<th>SNP5</th>
<th>SNP6</th>
<th>SNP7</th>
<th>SNP8</th>
<th>SNP9</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREE1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TREE4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE5</td>
<td>1</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TREE7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TREE8</td>
<td>0</td>
<td>NaN</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TREE10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
2. Filter genetic data: missing individuals
2. Filter genetic data: by allele frequencies

Minor Allele Frequency across all individuals

Major Genotype Frequency across all individuals

G. H. Hardy
2. Filter genetic data

- Genotypic data
- Environmental data

Sambada

Correlations
3. Evaluate spatial genetic variation – heterozygosity
4. Population structure

• Two options
  • Few/no missing values – principal component analysis, after filtering missing values
  • Without filtering – eigenvectors
4. Population structure
4. Population structure

Variation along first eigenvector
5. Summarize environmental data
5. Summarize environmental data

Variation along along PC1 across study area
5. Summarize environmental data

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PopStr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREE1</td>
<td>-8.648228102</td>
<td>3.173037266</td>
<td>3.66273579</td>
<td>-0.177338137</td>
<td>-8.554180345</td>
</tr>
<tr>
<td>TREE2</td>
<td>-8.648228102</td>
<td>3.173037266</td>
<td>3.66273579</td>
<td>-0.177338137</td>
<td>-12.43752863</td>
</tr>
<tr>
<td>TREE3</td>
<td>-8.648228102</td>
<td>3.173037266</td>
<td>3.66273579</td>
<td>-0.177338137</td>
<td>-7.394101741</td>
</tr>
<tr>
<td>TREE4</td>
<td>-8.570749832</td>
<td>2.761219441</td>
<td>4.189358619</td>
<td>0.120877968</td>
<td>-10.4768156</td>
</tr>
<tr>
<td>TREE5</td>
<td>-8.570749832</td>
<td>2.761219441</td>
<td>4.189358619</td>
<td>0.120877968</td>
<td>-8.093398487</td>
</tr>
<tr>
<td>TREE6</td>
<td>-8.570749832</td>
<td>2.761219441</td>
<td>4.189358619</td>
<td>0.120877968</td>
<td>-10.04374811</td>
</tr>
<tr>
<td>TREE7</td>
<td>-8.570749832</td>
<td>2.761219441</td>
<td>4.189358619</td>
<td>0.120877968</td>
<td>-10.37830172</td>
</tr>
<tr>
<td>TREE8</td>
<td>-8.570749832</td>
<td>2.761219441</td>
<td>4.189358619</td>
<td>0.120877968</td>
<td>-9.161504383</td>
</tr>
<tr>
<td>TREE9</td>
<td>-8.570749832</td>
<td>2.761219441</td>
<td>4.189358619</td>
<td>0.120877968</td>
<td>-9.732114439</td>
</tr>
<tr>
<td>TREE10</td>
<td>-8.570749832</td>
<td>2.761219441</td>
<td>4.189358619</td>
<td>0.120877968</td>
<td>-8.2234022058</td>
</tr>
</tbody>
</table>
6. Analyze with Sambada

Genotypic data

Sambada

Ranked markers

Environmental data
7. Display results – on map
7. Display results – in JBrowse
# Timeline

<table>
<thead>
<tr>
<th></th>
<th>Jul 17</th>
<th>Jul 31</th>
<th>Aug 10</th>
<th>Aug 31</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Filter trees</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Check compatibility</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Filter genetic data</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Spatial genetic variation</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Population structure</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Summarize environmental data</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Analyze with Sambada</td>
<td>✗</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Display results</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Documentation + tutorials</td>
<td></td>
<td>✗</td>
<td></td>
</tr>
</tbody>
</table>
Plant Computational Genomics Lab
Ecology and Evolutionary Biology
University of Connecticut
• Taylor Falk
• Emily Grau
• Jill Wegrzyn

Washington State University
• Stephen Ficklin
• Doreen Main

University of Tennessee, Knoxville
• Margaret Staton

Clemson University
• Alex Feltus

Semantic Options, LLC
• Damian Gessler

https://cartogratree.org
https://github.com/nicherndon/cartogratree

NSF National Science Foundation Awards DBI-0735191, DBI-1265383, and ACI-1443040

http://cartogratree.org  @treegenesdb@gmail.com  @treegenes  TreeGenes database