Case Study: TreeGenes uses SSWAP to Serve Semantic Data to Existing Web Applications, New Web Applications, and Semantic Pipelines

Presented by: Hans Vasquez-Gross
John Yu, Ben Figueroa, Damian Gessler, David Neale, Jill Wegrzyn

TreeGenes Database
Department of Plant Sciences
University of California at Davis
SSWAP Talk

Overview

• Who, Why, and What summary of the project
• TreeGenes & DiversiTree Introduction
• SSWAP Services
  – Built SSWAP Services
• SSWAP Integration within existing application
  – DiversiTree
• SSWAP Integration within new application
  – CartograTree
iPlant Seed Project: Tree Biology Cyberinfrastructure

- **WHO?**: Representatives from the Forest Tree Biology Community: Tree Physiologists, Forest Ecologists, Geneticists, Bioinformaticians

- **WHY?**: Physiological, population genetic, and ecological data lack shared, integrated, geo-referenced data storage

- **HOW?**: CyberInfrastructure to allow data storage, retrieval, integration, and analysis
  - Map-based tool to visualize the geo-referenced data

- **WHAT?** Genetic and genomic data from resequencing, transcriptome, and SNP studies associated with individuals with GPS coordinates. Phenotypes and environmental data with GPS.

- **Database Resources**:
  - Ameriflux (Abiotic Site Data in N. America), WorldClim (Abiotic Data Globally), Try-DB (Biotic/Abiotic Trait Data), TreeGenes (Genomics Data)
Nine modules to store and interrelate data for query and analysis in PostgreSQL

- Direct resource for nearly 2,500 forest geneticists representing 800 organizations worldwide. Over 6,000 unique visitors in December 2012.

- Forest Geneticists Colleague module
- Literature module
- Comparative map module
- Species module
- Sequencing/Primers module
- Genotype/SNP module
- Gbrowse/Genomic
- Phenotype/Expression module
- Sample tracking module
DiversiTree Resource
Advanced Workspace to Facilitate Search and Analysis of Resequencing Data

- Allows users to search from a variety of starting points
  - Functional Annotations (Candidate Gene)
  - Species
  - Genbank
  - ESTs
  - Sample/DNA Extraction Data
- Create customized lists, download data, and view in a variety of formats
  - FASTA format
  - Chromatogram Viewer
  - ACE Format
- Supports creation of custom SNP lists
  - Raw and normalized SNP scores
  - Genotype, flanking sequence, and alignments
  - Export to Genotyping formats (i.e. Illumina)
**DiversiTree Resource**

**Advanced Workspace to Facilitate Search and Analysis of Resequencing Data**
SSWAP Services

What is available

• To accomplish the project goals, web services need to be built to serve the relevant data

• Most services operate on a unique identifier, lending itself well to being integrated into existing applications
  – Ameriflux, Amplicon, Alignment, Contig, Contig keyword, EST, Multi-Fasta, Primer, Tree Sample, SNP, and Region sample

• These services have been integrated within DiversiTree by simply allowing users to select records then click a button to bring the data to SSWAP services.
<table>
<thead>
<tr>
<th>Service Name</th>
<th>Input</th>
<th>Description</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ameriflux</td>
<td>Ameriflux Site ID</td>
<td>Ameriflux site unique identifier (US-SP1)</td>
<td>RDF/XML: Elevation (m), Longitude (decimal degrees), Latitude (decimal degrees), Site Name (Ameriflux), SiteByName (Ameriflux name with state code), Site ID (Ameriflux), State, Country</td>
</tr>
<tr>
<td>Amplicon</td>
<td>Amplicon ID</td>
<td>Amplicon unique identifier (0_8156_01)</td>
<td>RDF/XML: SNP IDs, Alignment IDs, Forward and Reverse primer IDs</td>
</tr>
<tr>
<td>Alignment</td>
<td>Alignment ID</td>
<td>Alignment unique identifier (0_17973_01 20080627 Piel)</td>
<td>RDF/XML: Tree Sample IDs, SNP IDs, Taxa Record (NCBI hyperlink), Sequence</td>
</tr>
<tr>
<td>Contig</td>
<td>Contig ID</td>
<td>Contig unique identifier (0_10022)</td>
<td>RDF/XML: GO Term IDs, Contig name, Hyperlink to Contig record, FASTA sequence, Amplicon ID, Top BLAST hit, and EST name</td>
</tr>
<tr>
<td>Contig Keyword</td>
<td>GO Term Keyword</td>
<td>GO term description keyword (Heat Shock)</td>
<td>RDF/XML: Associated Contig IDs</td>
</tr>
<tr>
<td>EST</td>
<td>EST Name</td>
<td>Genbank/TreeGenes EST name (NXCI_123_A01_F)</td>
<td>RDF/XML: Contig ID, Library, Genbank Accession Number, Taxa Record (NCBI hyperlink), FASTA Sequence, and NCBI hyperlink to EST</td>
</tr>
<tr>
<td>MultiFasta</td>
<td>Amplicon ID</td>
<td>Amplicon unique identifier (0_8156_01)</td>
<td>RDF/XML: Multi-FASTA hyperlink</td>
</tr>
<tr>
<td>Primer</td>
<td>Primer ID</td>
<td>Primer set unique identifier (0_8156_01_R / 0_8156_01_F)</td>
<td>RDF/XML: Forward and Reverse primer IDs, Amplicon ID, Annealing temperature (degrees Celcius), and Sequence</td>
</tr>
<tr>
<td>Tree Sample</td>
<td>Tree Sample ID</td>
<td>Tree sample unique identifier (LADE0725)</td>
<td>RDF/XML: Elevation (m), Taxa Record (NCBI hyperlink), Longitude (decimal degrees), Latitude (decimal degrees), isGenotyped (yes/no), CartograTree hyperlink</td>
</tr>
<tr>
<td>Region Sample</td>
<td>Longitude, Latitude</td>
<td>Longitude and Latitude (decimal degrees)</td>
<td>RDF/XML: Site ID (Ameriflux), Site name (Ameriflux), SiteByName (Ameriflux name with state code), Distance (m), Longitude (decimal degrees), Latitude (decimal degrees), and Country</td>
</tr>
<tr>
<td>Region Sample</td>
<td>Distance</td>
<td>Distance (km)</td>
<td></td>
</tr>
<tr>
<td>SNP</td>
<td>SNP ID</td>
<td>SNP unique identifier (SNP_16271-Pisy)</td>
<td>RDF/XML: Taxa record (NCBI hyperlink), Polybayes score, Alignment ID, Sequence, Genotype, Amplicon ID, PolyPhred score, Relative position, Flanking sequence (1 and 2), Polymorphism length, and isGenotyped (yes/no)</td>
</tr>
</tbody>
</table>
DiversiTree

SSWAP Integration

Selected contigs of interest

Send data to sswap.info
DiversiTree

SSWAP Integration

Built SSWAP pipeline

Display Data Link to CartograTree
DiversiTree

SSWAP Integration

Resulting Visualization
CartograTree

Opening: Layout

Search and Filter Panel

[-] Search
Search for ID

[-] Display
Taxa: All

Tree Samples
Sequenced: [ ] Yes [ ] No
Genotyped: [ ] Yes [ ] No
Phenotyped: [ ] Yes [ ] No

Phenotype Search: Search traits here

GPS Resolution: [ ] Coordinates [ ] Approximation

TRY-DB Markers
TRY-DB: [ ]

Environmental
[ ] Ameriflux Sites

[+] Legend

[ - ] Selection Tools
Delete Selected Shape

[ - ] Layers
Repository
Layer: Select Layer——  Layer Info

Selected Markers

Menu

Proceed

Map data ©2013 Google, INEGI, MapLink, Tela Atlas, Terms of Use
CartograTree

Opening: Layout
CartograTree

Opening: Layout

Selected Markers Tools Panel
**TreeGenes Tree Sample**
901,113 sequences
24,142,786 genotypes
19,441 phenotypes

**TRY-DB Trait**
80,366 phenotypes
368 species
**AmeriFlux**
4,981 records
81 active remote sensing sites

**WorldClim**
50 years of dense, accurate recordings world-wide (interpolated)

**CartograTree**
Map Markers
CartograTree

Filter: By Taxa

Pinus Filter Applied
CartograTree

Filter: By Genetic Attribute

Phenotype Filter Applied
CartograTree
Filter: By Phenotype Term

Phenotype Term Filter Applied
CartograTree

Base Maps
Once the individual trees of interest are selected, a variety of tools are available including:

- **Common SNP, Common Phenotype, Get WorldClim Data** allow users to setup association studies
- **DiversiTree Input File** allows users to easily explore the data in DiversiTree
- **Common Amplicon** can be used to get sequence data and run a Multiple Sequence Alignment
- **SSWAP** brings the data to the SSWAP pipeline
# CartograTree

## “Common” DataSets

### Selected Markers

<table>
<thead>
<tr>
<th>Type</th>
<th>ID</th>
<th>Lat</th>
<th>Lng</th>
<th>Sequenced</th>
<th>Genotyped</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NC000870</td>
<td>32.972</td>
<td>-87.933</td>
<td>No</td>
<td>Yes</td>
<td>taeda</td>
</tr>
<tr>
<td></td>
<td>NC001209</td>
<td>34.171</td>
<td>-87.341</td>
<td>No</td>
<td>Yes</td>
<td>taeda</td>
</tr>
<tr>
<td></td>
<td>NC002049</td>
<td>36.435</td>
<td>-78.020</td>
<td>No</td>
<td>Yes</td>
<td>taeda</td>
</tr>
</tbody>
</table>

### Selected Markers

#### Common SNP

<table>
<thead>
<tr>
<th>TreeID</th>
<th>SNP_136591-P</th>
<th>SNP_136593-P</th>
<th>SNP_136598-P</th>
<th>SNP_136599-P</th>
<th>SNP_136600-P</th>
<th>SNP_136602-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC000182</td>
<td>AA</td>
<td>GG</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>NC001209</td>
<td>AG</td>
<td>AG</td>
<td>AG</td>
<td>AA</td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>NC002049</td>
<td>AG</td>
<td>GG</td>
<td>--</td>
<td>AG</td>
<td>AA</td>
<td>AA</td>
</tr>
</tbody>
</table>

### Selected Markers

#### Common Phenotypes

<table>
<thead>
<tr>
<th>TreeID</th>
<th>accuracy of breeding value for height</th>
<th>accuracy of breeding value for stem form</th>
<th>accs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC000182</td>
<td>0.600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NC001209</td>
<td>0.500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NC002049</td>
<td>0.630</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Selected Markers

#### Common Phenotypes

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Latitude</th>
<th>Longitude</th>
<th>TempAnomal</th>
<th>AnomTempRange</th>
<th>MaxTempWarmMonth</th>
<th>MinTempColdMonth</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC001209</td>
<td>34.17908</td>
<td>-87.24130000000002</td>
<td>15.5</td>
<td>34.2</td>
<td>32</td>
<td>-2</td>
</tr>
<tr>
<td>NC000182</td>
<td>29.275815</td>
<td>-82.12280299999998</td>
<td>20.8</td>
<td>26.6</td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td>NC002049</td>
<td>36.435213</td>
<td>-78.61959300000001</td>
<td>14.5</td>
<td>34.3</td>
<td>31.5</td>
<td>-2</td>
</tr>
</tbody>
</table>
• All three files can be downloaded (CSV) format for input into R or TASSEL.
• For the **Common Amplicon** example, the resulting amplicon data can be downloaded as a CSV or sent to SSWAP to create a pipeline.
• If sent to SSWAP, the **MultiFasta service** can be used to grab the actual sequence for the amplicons.
• This data can then be sent to a Multiple Sequence Aligner, such as MUSCLE
• The output could then be sent to FastTree and TreeViz for visualization as a phylogenetic tree
CartograTree

Phylogenetics

sswap.info beta
Simple Semantic Web Architecture and Protocol

New pipeline

Input Data Set

Display Data: Click here to send the output data to a viewer (opens in a new window)
Data Set URL: http://sswap.iplantcollaborative.org/lpc/rng?token=e4bebf6b-756f-41a8-a67c-618c630452e
Produced by: http://sswap.dendrome.ucdavis.edu/resources/ampliconService/ampliconService
Acknowledgements

TreeGenes:
Jill Wegrzyn
John Yu
Ben Figueroa
David Neale
Hans Vasquez-Gross

iPlant Collaborative:
Damian Gessler
Martha Narro
Nirav Merchant

TRY-DB:
Jens Kattge
Gerhard Boenisch

WorldClim:
Robert Hijmans
Questions and Comments

http://dendrome.ucdavis.edu/cartogratree

Learn More:
Poster TreeGenes: P0948
Poster CartograTree: P0992
Follow us 🌳: @TreeGenes

Read More: