

Beyond Search and Display – Analyze Tripal Data with CartograTree

Sean Buehler

Plant and Animal Genome Conference

January 13, 2019

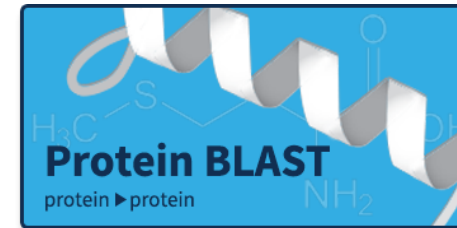


Tripal Possibilities

- Tripal is great for biological data!
 - Storage
 - Searching
 - Displaying
- Community Modules



CHADO





<https://treegenesdb.org>



- 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

A screenshot of the TreeGenes website homepage. The header features the TreeGenes logo and navigation links: TreeGenes, Community, Species, Literature, Tools, and Download Data. A large banner image of a forest with the text "Welcome to the new TreeGenes Database! Powered by Tripal!" is displayed. Below the banner is a section titled "EXPLORE TREEGENES" with six icons and links: Browse Genomes (Browse genomes with JBrowse), CartograTree (CartograTree is a map-based web app), Download Data (Download data from our FTP site), Sequence Search (Search sequences with DIAMOND), Species (Find your species of interest), and Submit (Submit your data to TreeGenes). To the right is a "USER MENU" with links: Edit Profile, Log out, My account, Add new job posting, and Add new meeting posting. Below the explore section is a "FEATURED PROJECTS" section with four small images of plants. On the far right, there is a "MEETINGS" section with the text "PAG XXVII January 12 to 16 2019 San Diego, USA".



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

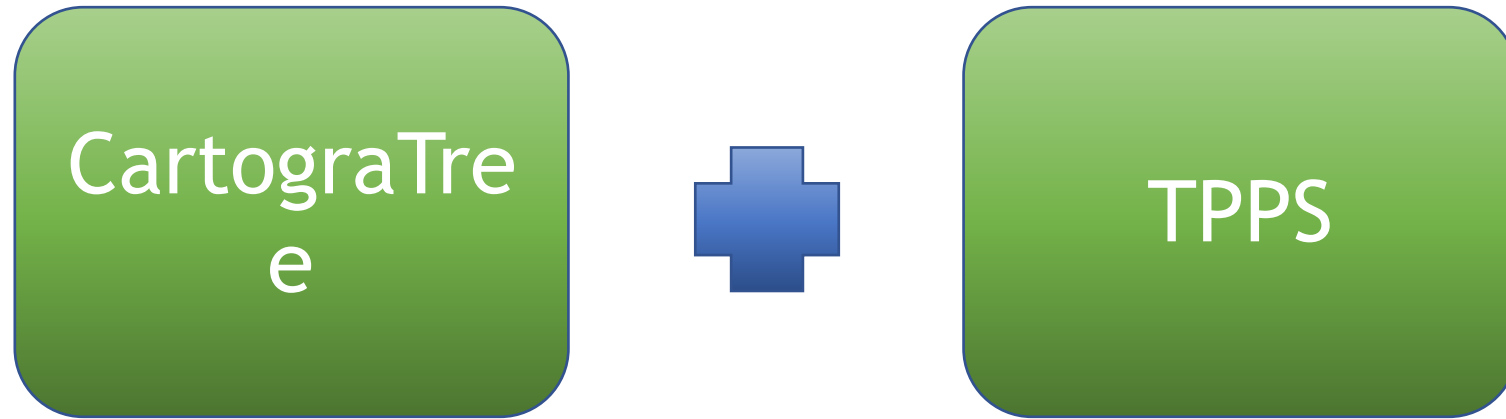


The current process

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

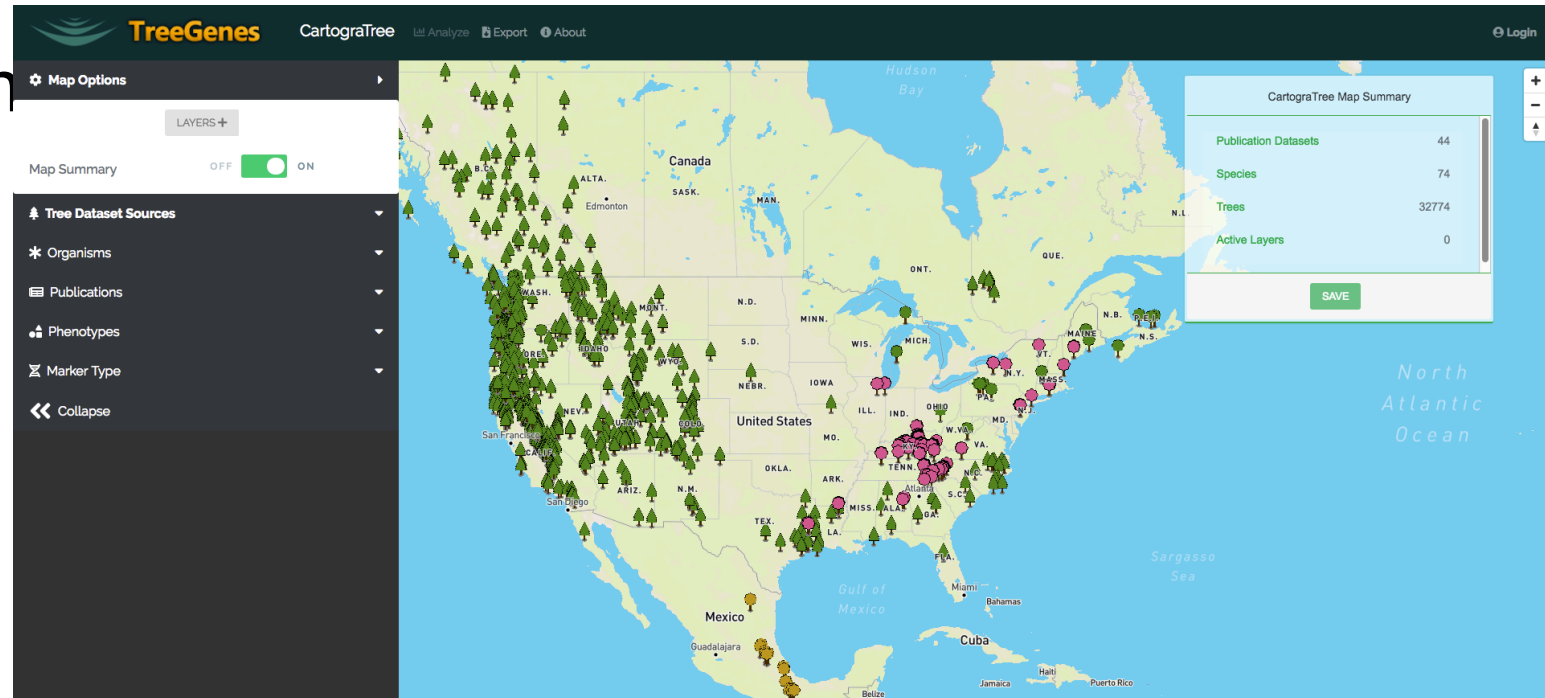


Solution?



CartograTree

- Web-based map-driven
 - Genotypic
 - Phenotypic
 - Environmental



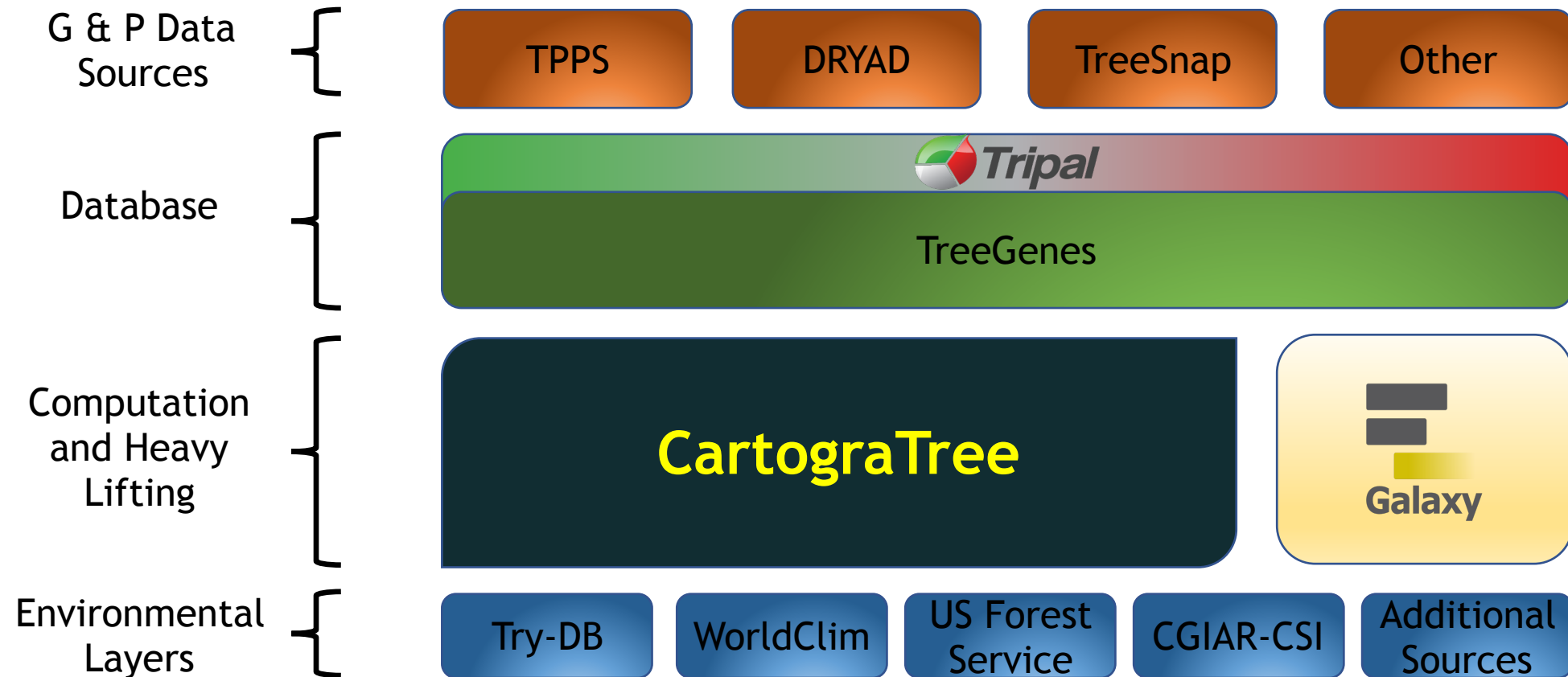
How it Works



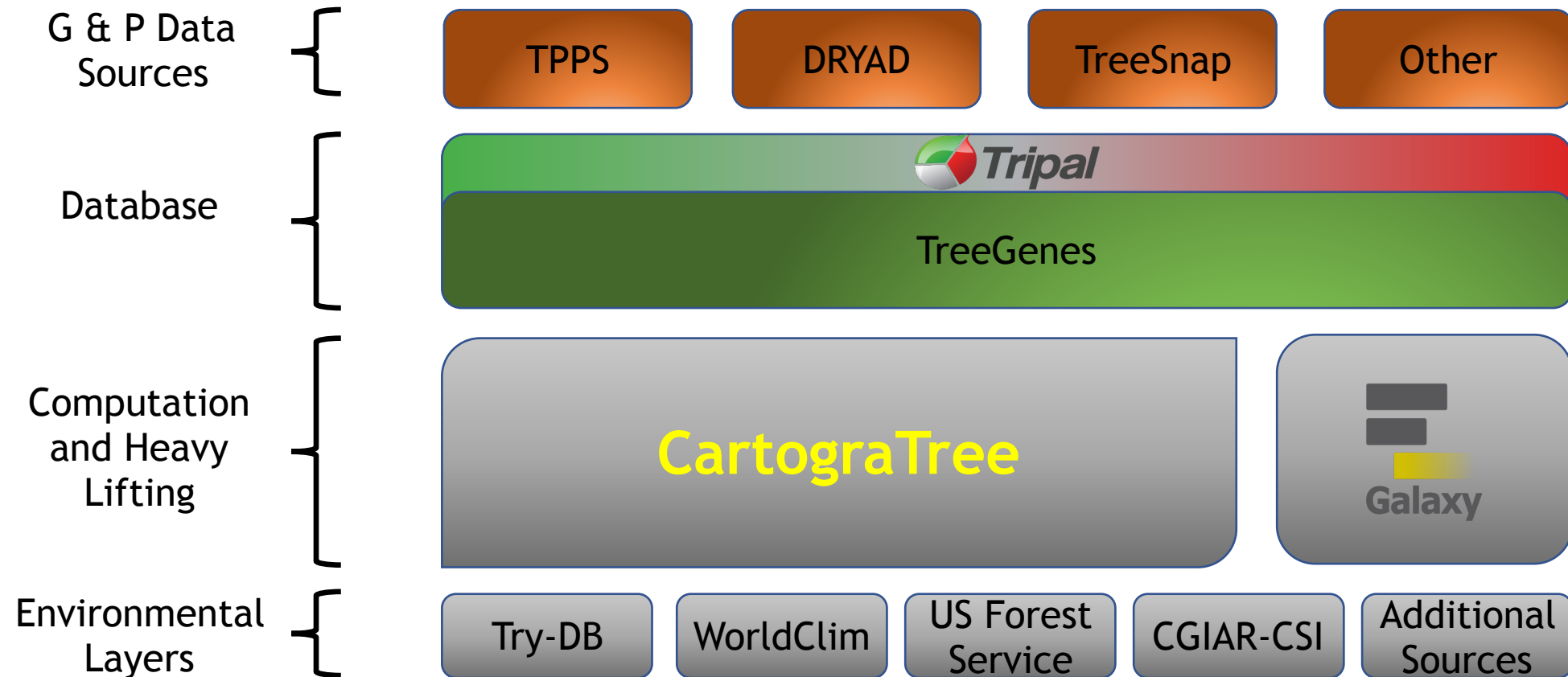
- CartograTree is a Tripal extension module
 - PHP
 - NodeJS - API
 - MapBox
 - Geoserver
- Access existing data from the Tripal site (CHADO)
- Access environmental data from public repositories
- Integrate data and perform analysis through Galaxy



General Overview



Clade/Model Organism Database



Tripal Plant PopGen Submit Pipeline

What it is

- Short yet comprehensive series of forms
- Form adapts to the user's input

Motivation

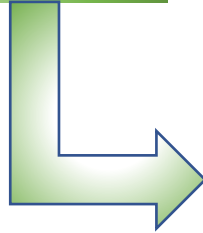
- Metadata surrounding a study:
 - Inconsistent
 - Sometimes lost after publication



Tripal Plant PopGen Submit

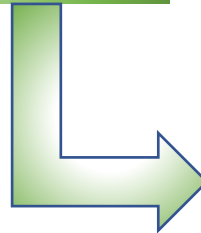
Population Study

- Publication
- Species



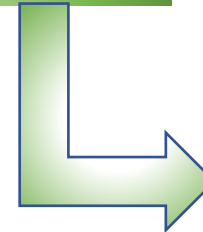
Study Design

- 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

Primary Author: *

Organization: *

▼ PUBLICATION INFORMATION:

ADD SECONDARY AUTHOR REMOVE SECONDARY AUTHOR

☐ I have >30 Secondary Authors

Publication Status: *
- Select -

Title of Publication: *

Abstract: *

Journal: *

▼ ORGANISM INFORMATION:

Up to 5 organisms per submission.

ADD ORGANISM REMOVE ORGANISM

Species 1: *

* : Required Field

SAVE NEXT

Data Type: *
Genotype x Phenotype x Environment

Study Type: *
Greenhouse

▼ GREENHOUSE INFORMATION:

Air Humidity controlled or uncontrolled: *
Uncontrolled

Light Intensity controlled or uncontrolled: *
Uncontrolled

TEMPERATURE INFORMATION:

Please provide temperatures in Degrees Celsius

Average High Temperature: *
40

Average Low Temperature: *
20

ROOTING INFORMATION:

Rooting Type: *
Aeroponics

pH controlled or uncontrolled: *
Controlled

Controlled pH Value: *
6.0

TREATMENTS: *

☐ Seasonal Environment

☐ Air temperature regime

☐ Soil Temperature regime

☐ Antibiotic regime

☐ Chemical administration

☐ Disease status

☐ Fertilizer regime

☐ Fungicide regime

☐ Gaseous regime

☐ Gravity Growth hormone regime

☐ Mechanical treatment

☐ Mineral nutrient regime

☐ Humidity regime

Author and Species Information Experimental Conditions Tree Accession Submit Data

Tree Accession File: please provide a spreadsheet with columns for the Tree ID and location of trees used in this study: *

☐ tpss_accession_test.xlsx REMOVE

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: Tree Identifier and Location

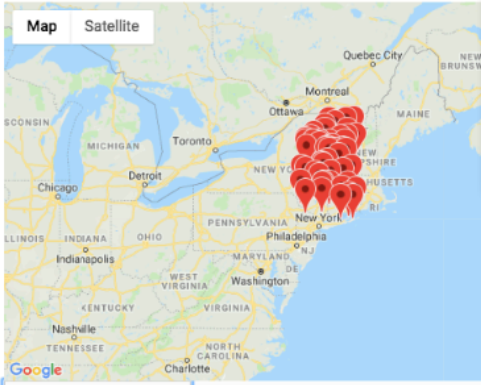
latitude	Longitude	tree id	Tree Identifier	extra
43.350835454785027	-72.091250275419483	tree1		
41.896758401561473	-74.129167101588678	tree2		
41.206891072226776	-74.6447138066612316	tree3		

☐ My file has no header row

Please upload a spreadsheet file containing tree population data. When your file is uploaded, your column header names, several drop-downs, and the first few rows of your file. You must provide a column for each column, using the drop-downs provided to you. If a column data type does not fit a menu, you may omit that drop-down menu. Your file must contain columns with information and the Location of the tree (either gps coordinates or country/state).

Coordinate Projection
WGS 84

Map Satellite



Click here to view trees on map!

Map data ©2018 Google, IN Terms of Use

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

☐ phenotype_metadata.xlsx REMOVE

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

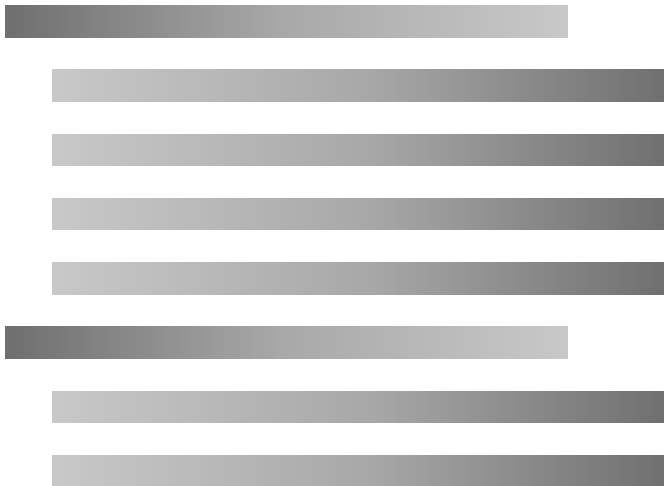
name	attribute	units	type
Phenotype Name/Identifier	Attribute	Units	Description
phenotype 1	age	years	quantitative
phenotype 2	age	years	quantitative
phenotype 3	age	years	quantitative





Minimum Information About a Plant Phenotyping Experiment

Study



TPPS



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


TreeGenesCommunitySpeciesLiterature

TPPS Details by Accession

[Return to TPPS List](#)

TPPS/TGDR DETAILS FOR TGDR008

Attribute	Details
Accession	TGDR008
Title	Molecular analysis of natural root grafting in jack pine (<i>Pinus banksiana</i>) trees: how genetic proximity influenced anastomosis presence?
Species	<i>Pinus banksiana</i>
Study Type	GxE
File Downloads	 GENOTYPES SSR  GPS COORDINATES
CartograTree	View in CartograTree
Tree Count	105

TreeGenesCartograTreeAnalyzeAboutLogin

Map Options


LAYERS +

Map Summary OFF ☒ ON

Tree Dataset Sources



- Organisms
- Publications
- Phenotypes
- Marker Type

<< Collapse



Latitude: 47.975 | Longitude: -77.388
Coordinate type: exact

Pinus banksiana (gymnosperm)
Family: Pinaceae

ID: TGDR008-100380  

1 of 105

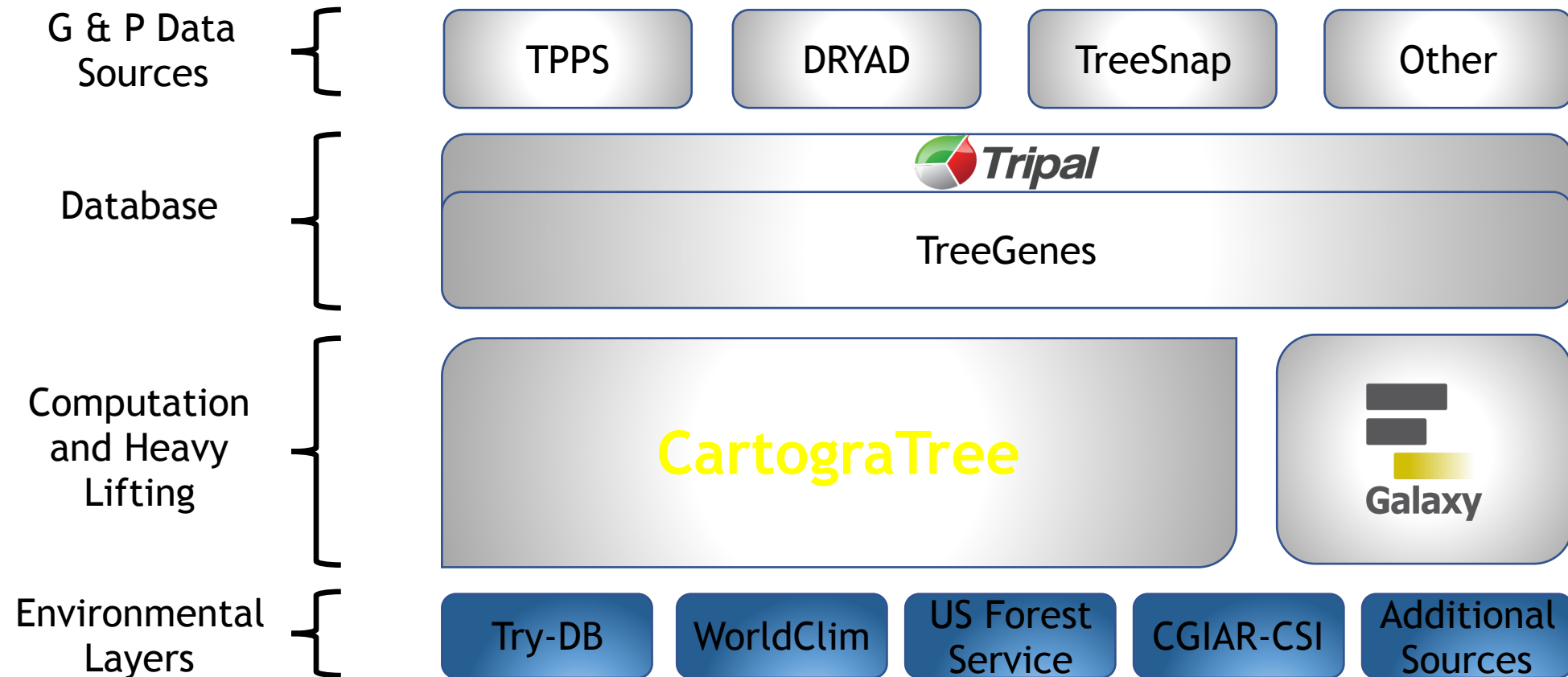
CartograTree Map Summary

Publication Datasets	46
Species	63
Trees	23860
Active Layers	0

SAVE



Environmental Layers



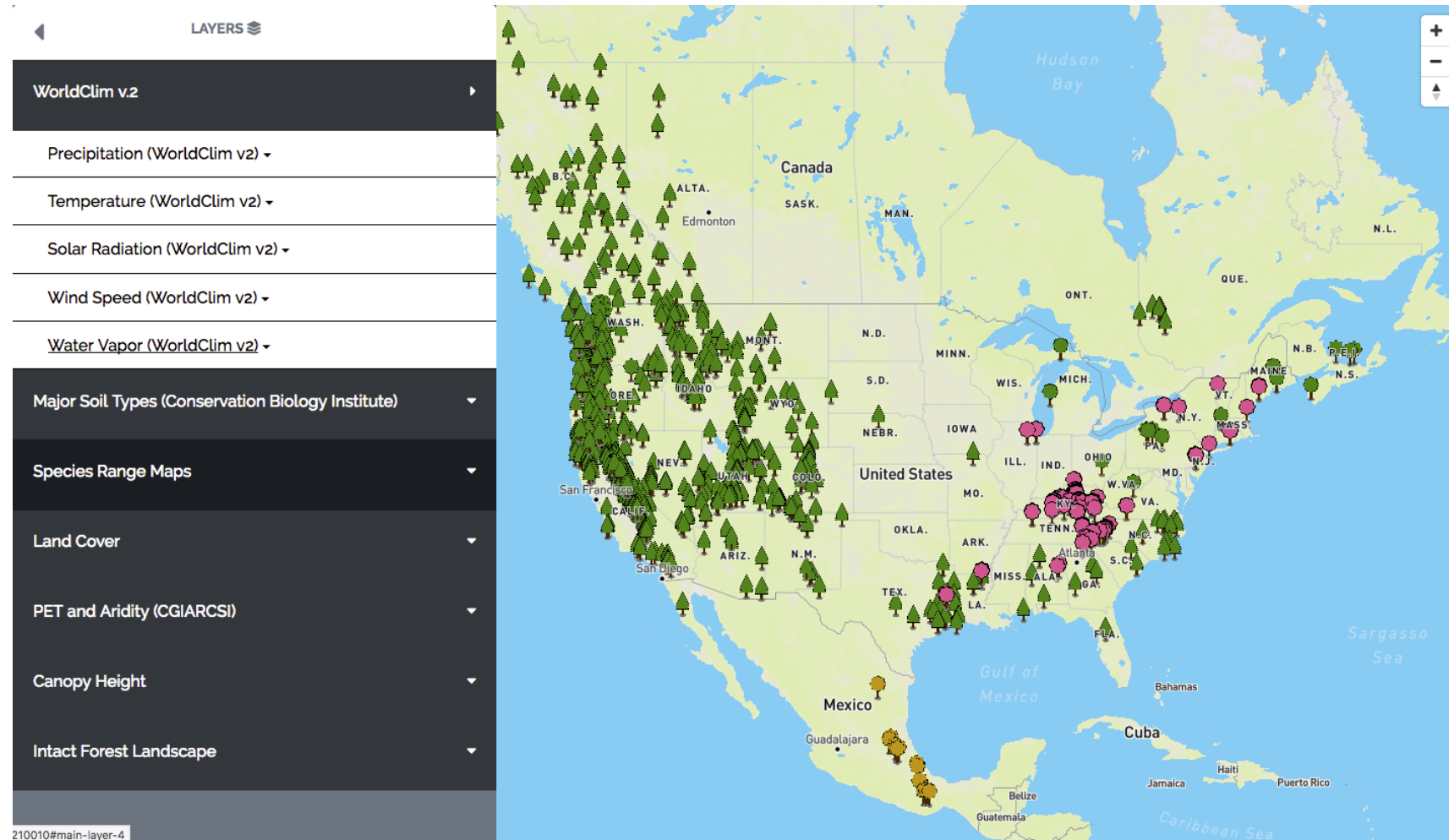
Environmental Layers

WorldClim – Global Climate Data

- 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

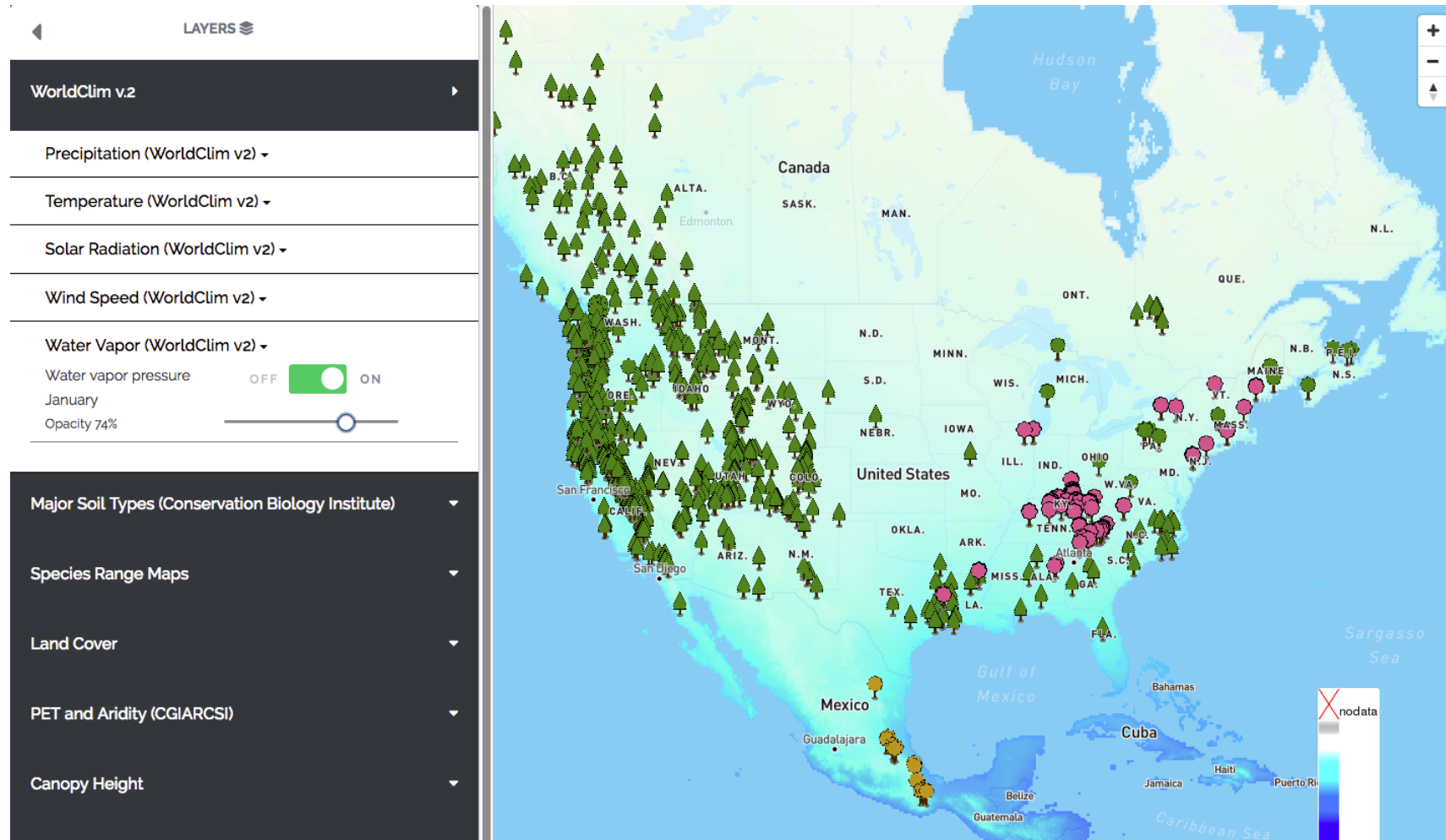


Environmental Layers



Environmental Layers

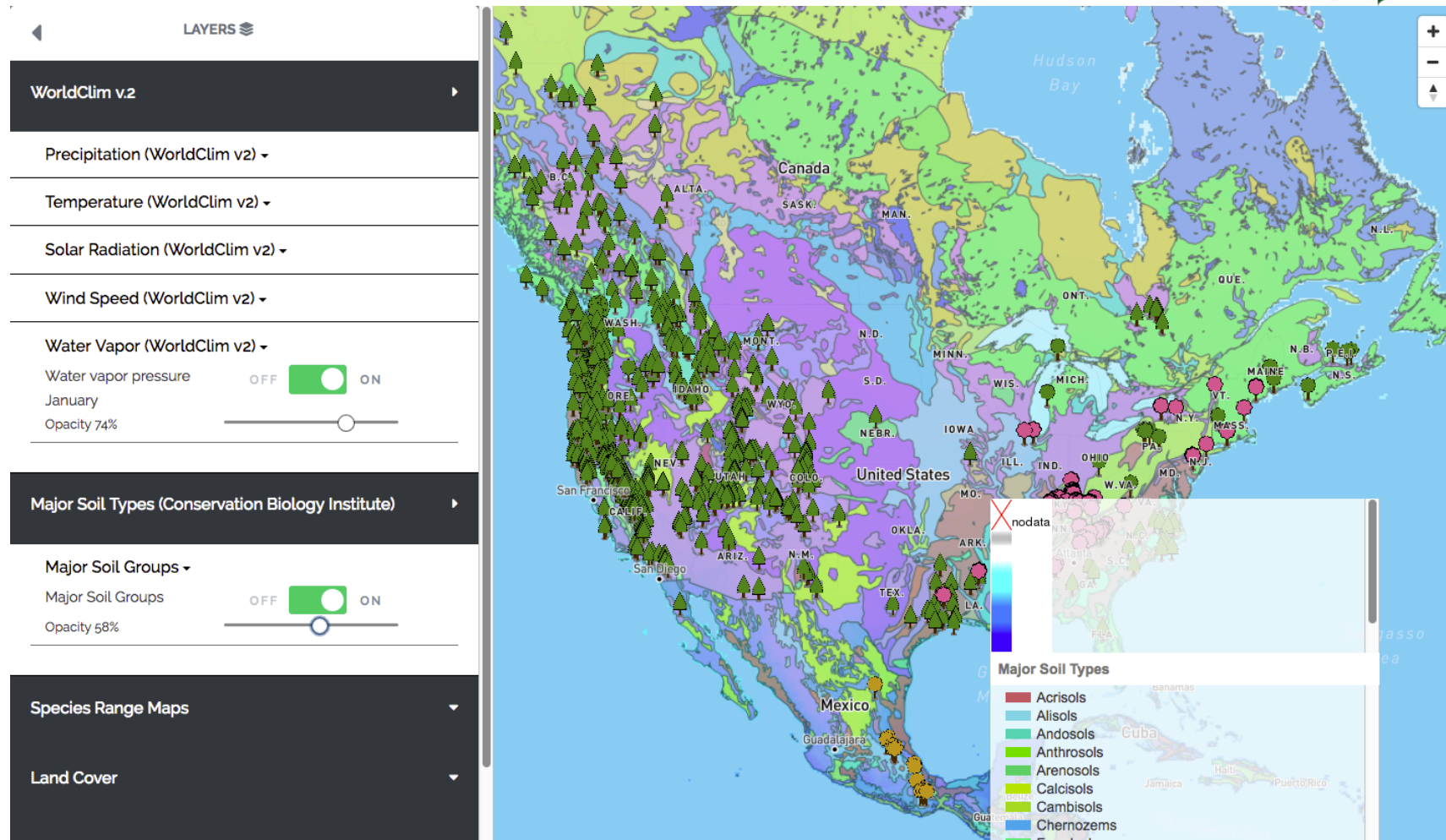
WorldClim – Global Climate Data



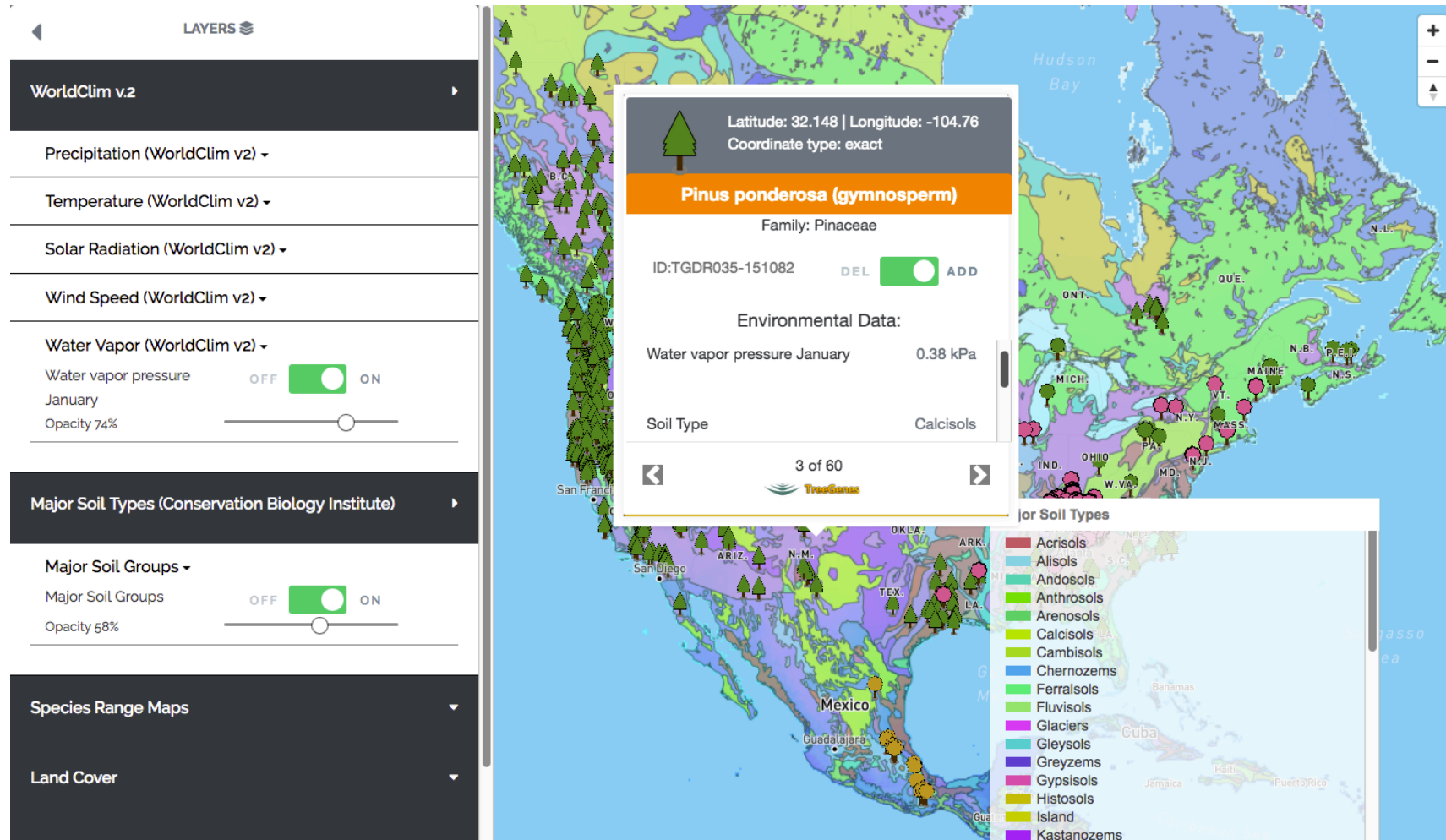
Environmental Layers



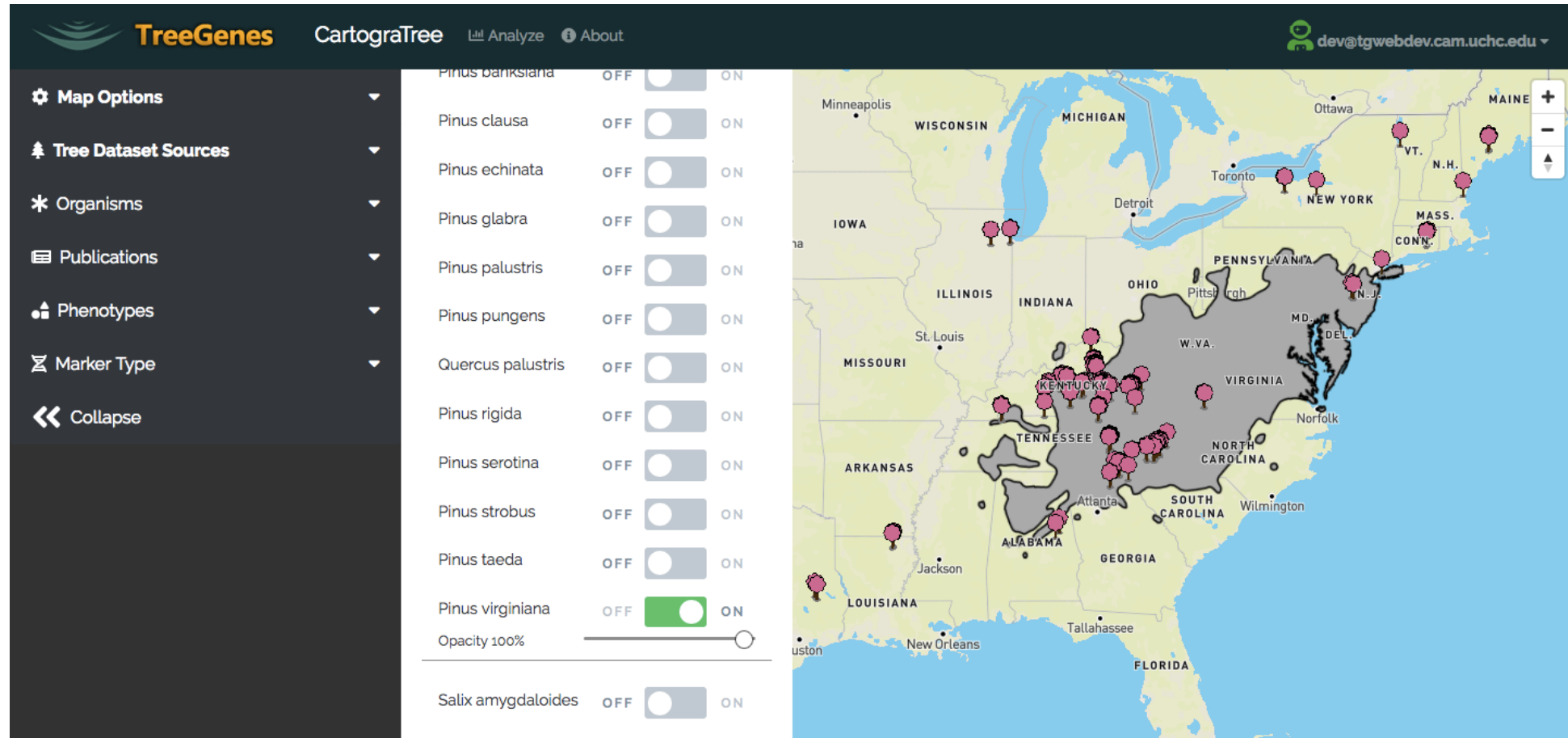
Conservation
Biology Institute



Environmental Layers



Species Range



CartograTree - Searches

Searches saved to User Profile

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

Select State of Map

Additional Options

Summary and Confirm

Current Map State

Select State of Map

Additional Options

Summary and Confirm

Analysis type

Landscape GxE ▾

Publications Selected

Environmental variables

BACK

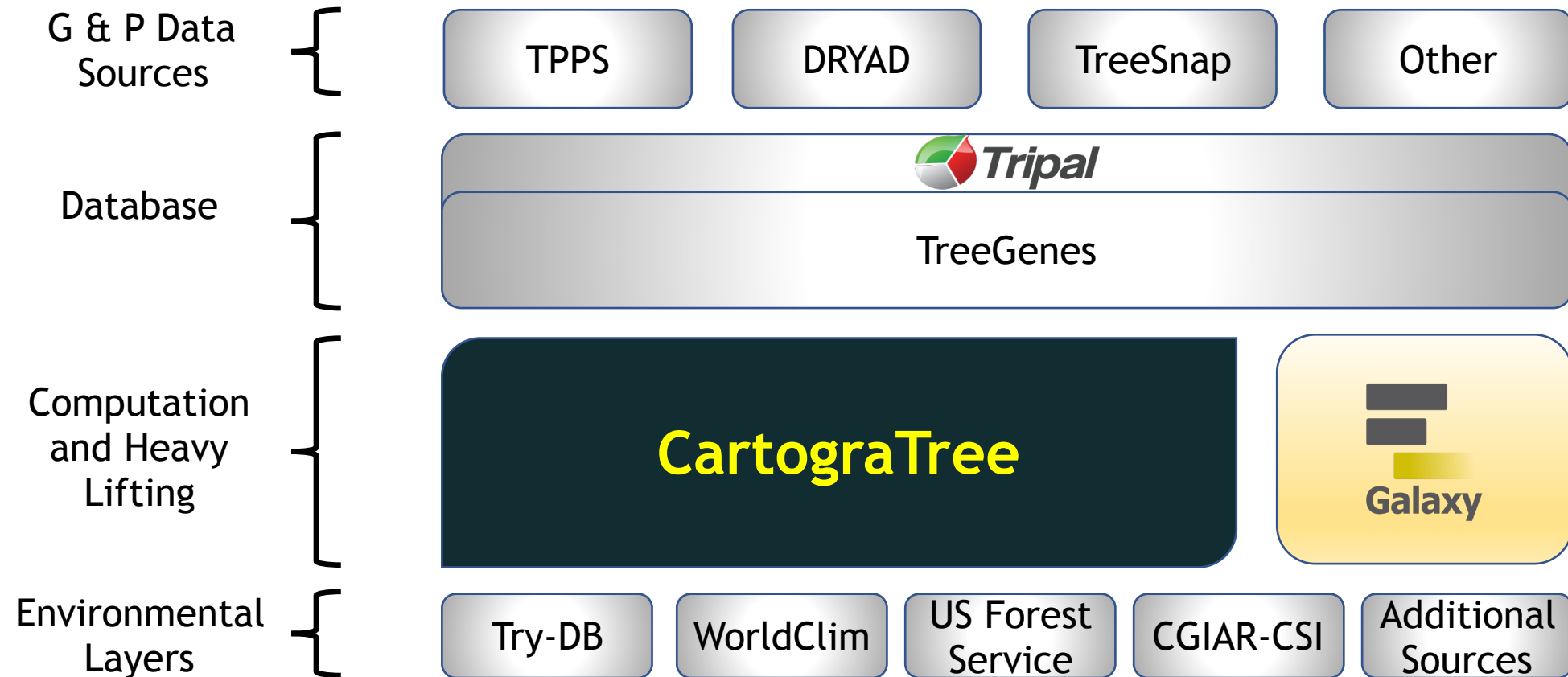
NEXT

LOAD

NEXT



Analysis Stage



Galaxy PROJECT



TriPal

Galaxy / TreeGenes Analyze Data Workflow Visualize Shared Data Admin Help User Using 2.3 GB

Tools Workflow Canvas | Sean Sambada

Filter and Sort

GFF

- Extract features from GFF data

Join, Subtract and Group

- Group data by a column and perform aggregate operation on other columns.

Convert Formatters

- Convert BED coverage

Phenotype Association

- aaChanges amino-acid changes caused by a set of SNPs
- LD linkage disequilibrium and tag SNPs
- BEAM significant single- and multi-locus SNP associations in case-control studies

AMTools

- LANDSCAPE GENOMICS WITH SAMBADA
- SamBada landscape genomic analysis (Galaxy Version 0.5.3)

Parameters to set up analysis

- Environmental data

univariate_model (txt)

bivariate_model (txt)

log (txt)

Label

Add a step label.

Annotation

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

Parameters to set up analysis

Data input 'params' (txt)

Environmental data

Data input 'env_input' (txt)

Molecular data

Data input 'mol_input' (txt)

Email notification

TriPal Galaxy API

<https://galaxyproject.org>



Preparing data for analysis

1. Filter chosen trees/ data

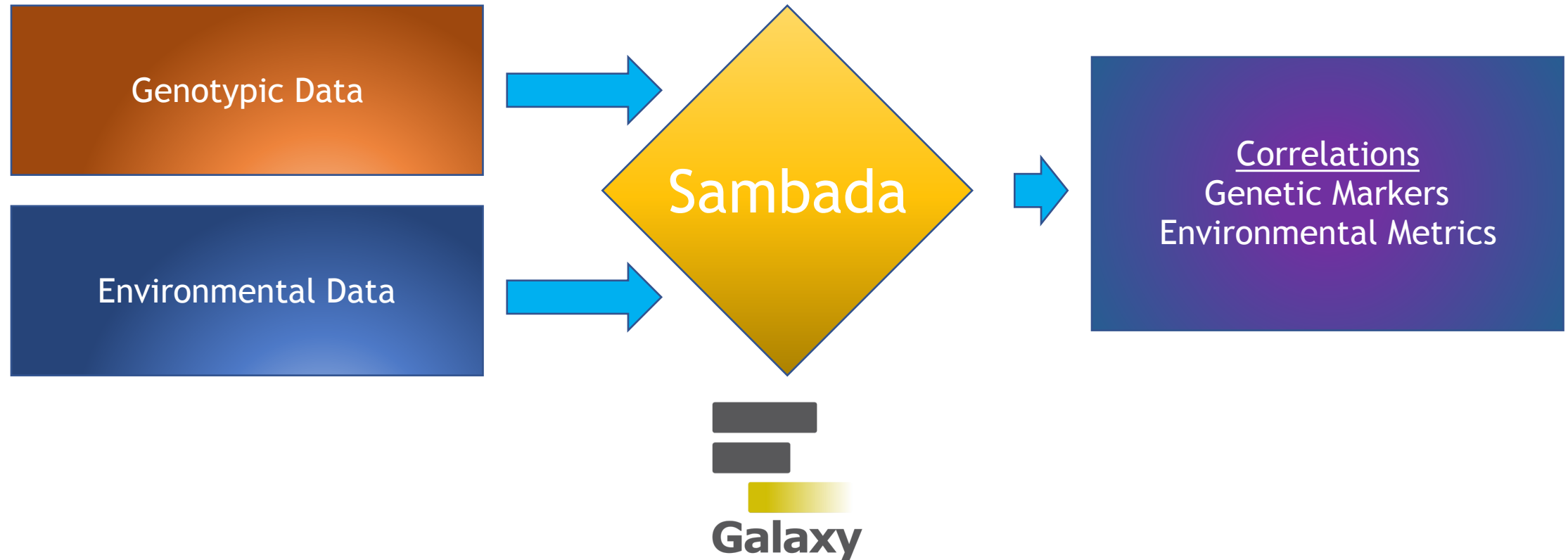
- Missing data

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SN
TREE1	0	0	0	0	0	0	0	0	0	
TREE2	1	0	2	0	0	0	0	0	0	
TREE3	1	1	0	0	0	0	0	1	1	
TREE4	1	1	0	0	0	0	0	0	0	
TREE5	1	0	NaN	0	0	1	0	0	0	
TREE6	1	0	1	0	0	0	0	1	0	
TREE7	0	1	1	0	0	0	0	1	1	
TREE8	0	NaN	0	0	0	1	0	0	1	
TREE9	1	0	1	1	0	1	0	0	1	

2. Choose appropriate workflow from Galaxy



Example workflow: Landscape Genomics



Conclusion

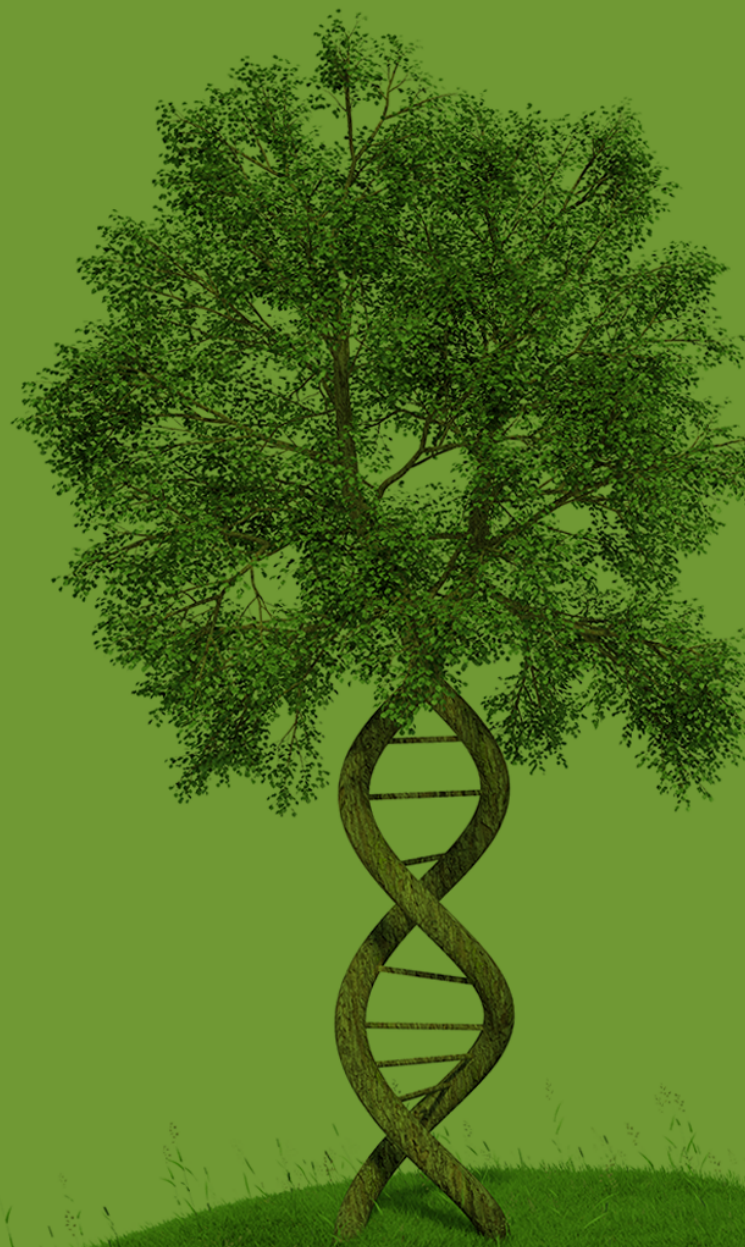
- Building on the functionality of Tripal
 - Phenotypic, genotypic, environmental + metadata
 - community modules
- TPPS - A pipeline to collect standardized and necessary data
- CartograTree - A web-based tool to integrate data and launch analyses



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: P00527

@TreeGenes 



Awards
DBI-0735191,
DBI-1265383,
ACI-1443040



Conservation
Biology Institute



WorldClim – Global Climate Data

