What is Tripal?

Tripal (https://tripal.info) is the software framework used by TreeGenes. It is freely-available and open-source software that enables the creation of modules for the Drupal content management system that are ideal for storing, displaying, and analyzing genetic and phenotypic data from both model and non-model systems. All data and its relationships that are housed within the underlying Chaos databases are made available within the context of controlled vocabulary terms. Tripal is also highly customizable, since it is open-source and provides its own API, which enables anyone to create their own tools and extensions.

What is CartograTree?

CartograTree is a map-based interface for viewing the connection between genotypic and phenotypic data associated with georeferenced trees from the TreeGenes database and environmental metrics from sources such as TopoDB, AmeriFlux, and WorldClim, thereby enabling association analysis. Once data has been selected for analysis, it is transferred to a high-performance computing system running Galaxy, where it is analyzed using Galaxy workflows built specifically for CartograTree. Current development is focused on supporting complex data selection across multiple environmental layers simultaneously.

General Workflow for TPPS

Users will provide files with information detailing the location of the trees when they were collected. (Figure 3C) Finally, users will fit out forms and upload files with genotypic, phenotypic, and environmental data. (Figure 3D) At the end of each step, data is validated to ensure that it is usable and complete. (Figure 3E) For example, when the user submits their genotypic data, the tree identification numbers are compared with the tree identification numbers of the file that was provided in the Tree Access step. This verifies that all of the user’s genotypic data can be mapped to a tree with a known geographic location.

What is TreeGenes?

TreeGenes is a web-based repository, which provides custom informatics tools to manage the flood of information resulting from high-throughput genomics and phenomics projects in forest trees from sample collection to downstream analysis. The database itself contains a variety of genetic data types, including transcriptomes, expression studies, variant data, genetic maps, and full genome for over 1700 forest tree species. The database also contains phenotypic information and environmental data in relation to georeferenced trees. TreeGenes is enhanced with systems that are well connected with federated databases, automated data flows, standardized annotations and quality control processes. The recent conversion of TreeGenes into Tripal allows for features such as coarse-graining and the ability to move large data sets into analytical frameworks hosted on high-performance computing (HPC) systems. Association genetics can be used to identify traits that are likely to have certain traits, including but not limited to: adaptation to a changing climate, disease resistance, and wood quality/growth traits. Genomic-based genetic analysis can be used to determine how trees respond to environmental factors, which can be important to determining whether a tree population will persist despite climate change. (Figure 1)

What is TPPS, and how is it relevant to Galaxy?

The Tripal Plant PhyGen Submit pipeline (TPPS) is an open-source extension module for Tripal, which seeks to streamline the collection, organization, ontology assignment, and integration of Genotype, Phenotype, and Environmental data and metadata. (Figure 1) Having the combination of this data in one centralized database allows for analyses that are not feasible, or are difficult to conduct with only Genotype data, such as population genetics, ecological genetics, or landscape genomics. TPPS collects, organizes, and integrates raw data from users so that it can be properly used by Galaxy tools. A general flowchart for the intended data flow from raw data to a visual map analysis can be seen in Figure 2.

Plant Phenotyping Standards

MAPPE, the Minimum Information About a Plant Phenotyping Experiment, is a set of attributes that may be necessary to fully describe a plant phenotyping experiment. One of our goals was to create a pipeline that takes in information in accordance with these standards. Requirements can be broken down into the following categories:

- General metadata
- Imaging, location, and Environment
- Biosource, Treatments, and Experimental design
- Sample Collection, Processing, and Management
- Observed Variables

The purpose of following these MAPPE standards is to ensure that the data we collect has a certain level of integrity, so that when it is passed to CHADO, Tripal, Galaxy, etc, it will not contain holes where there should have been information. Since these standards are also the minimum information, we can verify that we are not asking any information that would not be relevant to the study, thus providing a more streamlined submission process.

Data Integration Goals

Once a TPPS submission is complete and approved, georeferenced data submitted through the pipeline can then be accessed by tools like CartograTree, which can query environmental layers, measuring factors like precipitation, temperature, and elevation, using Galaxy workflows. Researchers can leverage local or distributed HPC resources such as the Texas Advanced Computing Center (TACC) to access tools like their information and it has been approved by an administrator on the TreeGenes site, they will receive a accession number which can be used to reference the data and metadata in publications.

ACKNOWLEDGEMENTS

Funding from the NSF DIBBS Award Numbers 1443040 and 1444573

GENERAL WORKFLOW FOR TPPS

Users will provide files with information detailing the location of the trees when they were accessed. Users will fit out forms and upload files with genotypic, phenotypic, and environmental data. At the end of each step, data is validated to ensure that it is usable and complete. For example, when the user submits their genotypic data, the tree identification numbers are compared with the tree identification numbers of the file that was provided in the Tree Access step. This verifies that all of the user’s genotypic data can be mapped to a tree with a known geographic location.

Once all of the data has been validated by the pipeline and reviewed by the submitter, it is input into the CHADO database schema, and placed in custom temporary tables, where it awaits approval from an administrator. The user receives an email that notifies them that the submission is pending approval, and a brief summary of the data provided.

To minimize the amount of data that must be stored locally, as well as the amount of data the user must input manually, we will ask for NCBI BioProject accession numbers, and allow for files rather than form fields whenever possible. After the user has provided all of their information and it has been approved by an administrator on the TreeGenes site, they will receive an accession number which can be used to reference the data and metadata in publications.