

TreeGenes and CartograTree: Enabling Visualization and Analysis in Forest Tree Genomics¹

E.S. Grau,² S.A. Demurjian,² H.A. Vasquez-Gross,³ D.G. Gessler,⁴ D.B. Neale,³ and J.L. Wegrzyn²

Association studies integrating environmental, phenotypic, and genetic data are key in understanding forest tree resilience to climate change and disease. As genomic resources increase, both in terms of complete reference sequences and magnitude of individuals genotyped, researchers are better equipped to identify correlations between genetic variation and adaptive or commercial traits. Computational resources designed to integrate and distribute high quality genotypic, phenotypic, and environmental datasets remain insufficient for the task.

TreeGenes is a web-based resource dedicated to the forest tree community. The database hosts genetic data from >1700 tree species, collected from primary databases such as GenBank, genomes and transcriptomes from dozens of species, genetic maps and phenotypic/association study data submitted by users. Additionally, TreeGenes hosts custom-developed tools that allow researchers to make the most of our genetic offerings. Through the website, users can access a custom Laboratory Information Management System (LIMS), download bulk datasets from the database, and visualize genetic and genomic data via GMOD tools (e.g. CMAP, GBrowse).

CartograTree is a web-based application hosted through TreeGenes that allows researchers to identify, filter, compare, and visualize geo-referenced biotic and abiotic data. Its goal is to support numerous multi-disciplinary endeavors including phylogenetics, population structure, and association studies. These goals are supported and enabled through Simple Semantic Web Architecture and Protocol (SSWAP), which leverages high-performance computing and data storage. Development on CartograTree will expand the available datasets and analytical capabilities. The map interface will include new layers such as forest fragmentation and climate shift predictions. TreeGenes' upcoming transition to Tripal, a Chado- and Drupal-based web content management system, will allow access to more data through connections to the Hardwood Genomics Project and Genome Database for Roseaceae, and to powerful analytical pipelines and computational resources through the Tripal module connection to the Galaxy Project.

¹ A version of this paper was presented at the Gene Conservation of Tree Species – Banking on the Future Workshop, May 16-19, 2016, Chicago, IL.

² University of Connecticut, 2131 Hillside Road, Storrs, CT 06269.

³ University of California Davis, One Shields Road, Davis, CA 95616.

⁴ Semantic Options, Santa Fe, NM 87508.

Corresponding author: emily.grau@uconn.edu.