

Community Standards for Genomic Resources, Genetic Conservation, and Data Integration¹

Jill Wegrzyn,² Meg Staton,³ Emily Grau,² Richard Cronn,⁴ and C. Dana Nelson⁵

Genetics and genomics are increasingly important in forestry management and conservation. Next generation sequencing can increase analytical power, but still relies on building on the structure of previously acquired data. Data standards and data sharing allow the community to maximize the analytical power of high throughput genomics data. The landscape of incomplete submissions, minimal metadata, and distributed raw data inhibits aspects of data re-use, validation, and discovery. Community databases can address these challenges for their research communities by collecting and integrating data, as well as providing analytical tools.

The Hardwood Genomics Web (HWG, <http://www.hardwoodgenomics.org>) and TreeGenes Database (<http://treegenesdb.org>) are community resources for forest tree genetics research. These web-based repositories host a variety of data types, including genome sequences, transcriptomes, genetic maps, and supporting marker information for 23 tree species in HWG and another 1780 species in TreeGenes. TreeGenes and the HWG host data visualization and analytical tools to further integrate and interrogate these datasets. Specific analytical functions include gene expression, functional annotation, and association genetics.

Recent developments in TreeGenes and HWG have focused on the organization and integration of phenotypic and environmental data for georeferenced tree accessions originating from studies around the world. The majority of phenotypic data is not collected in a form that can easily be integrated and generally resides in flat files or in disparate databases. Genomics is more centralized but the critical metadata components are often not required by primary repositories such as Genbank and therefore, never documented. Environmental data, in the form of layers, is distributed by numerous independent sources and there are few standards imposed to facilitate cross-query and integration. The tree databases have implemented web-based modules designed to capture data, metadata, and relevant accessions on all aspects of association genetics and population genomics studies. The acquisition of this data has made it possible to develop interfaces, such as CartograTree (<http://cartogratree.org>), which allows users to interact with map-based utilities to query georeferenced forest tree accessions, integrate complex environmental and phenotypic data, and perform association genetic studies through high-performance computing resources in the Cloud. HWG and TreeGenes are collaborating to offer increased connectivity and analytical capabilities through involvement with the Tripal project (<http://tripal.info>). Tripal is an open source software project aimed at streamlining the creation and maintenance of web-based community databases.

Efforts to expand offerings through HWG and TreeGenes are coordinated with the Forest Health Research and Education Center (FHREC, <http://www2.ca.uky.edu/Forestry/fhrec/index.html>). This center represents the intersection of participating U.S. Department of Agriculture Forest Service research stations, academic institutions, and related organizations. The FHREC aims to understand the genetics of pests and pathogens, contribute to technology development for tree improvement, support education and outreach activities, and coordinate collaboration between different players in forest tree conservation. The

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 Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs CT 06269.

³ Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, TN 37996.

⁴ USDA Forest Service, Pacific Northwest Research Station, 3200 SW Jefferson Way, Corvallis, OR 97331.

⁵ USDA Forest Service, Southern Station, Lexington KY 40546.

Corresponding author: jill.wegrzyn@uconn.edu.

goal to improve sustainability and address forest health concerns is met through coordinating outreach activities, information resources, and research objectives. Community databases collaborate with the FHREC by facilitating data and resource sharing with the scientific community. Collaboration among community databases and these organizations increases data availability, data integration, and data analysis capabilities.

The HWG and TreeGenes databases conducted a community survey to determine user priorities for data standards and sharing. The results formed the basis for a focused discussion at the ‘Community Standards for Genomic Resources, Genetic Conservation, and Data Integration’ session. The survey gathered 42 responses from academic, government, and private organizations. Users identified the integration of phenotypic, environmental, and genetic marker data as the highest priority. Only 30 percent of respondents directly interact with high performance computing (HPC) resources that might enable the analysis and integration of this data. The discussion supported the idea that community databases can better serve the forest tree research community by providing access to this integrated data as well as analysis tools directly linked to HPC resources. Discussion also focused on the implementation of data standards in terms of genetic marker types. This discussion also spoke to limitations of frequently used genetic and genomic resources for assessing forest populations. This was not limited to the data itself but rather encouraged discussion on what markers were most useful in a comparative genomics context. Experimental implementation of non-standardized markers produces results that are less informative and not cumulative. Parallels can be made to the human genetics community where standardized marker types have vastly improved the ability to build on studies on the biomedical field. Proper marker development requires a list of high quality gene sequences from the target species or a close relative (with less than 5 percent sequence divergence). Continued focus on developing high quality genomic resources (full transcriptome or genome) can achieve this.

Components of the workshop, survey, and discussion revealed that the community is concerned about data access, distribution, and integration. Model and clade organism databases remain critical to achieve this goal. The challenges lie in not just generating the appropriate genomic resources, but also in how to store this information so that it can promote access, validation, and discovery.