frontiers in Plant Science

MINI REVIEW
published: 25 June 2019
doi: 10.3389/fpls.2019.00813

Cyberinfrastructure to Improve Forest Health and Productivity: The Role of Tree Databases in Connecting Genomes, Phenomes, and the Environment

Jill L. Wegrzyn[1], Margaret A. Staton[2]*, Nathaniel R. Street[3]*, Dorrie Main[4]*, Emily Grau[1]*, Nic Herndon[1], Sean Buehler[1], Taylor Falk[1], Sumaira Zaman[1], Risharde Ramnath[1], Peter Richter[1], Lang Sun[1], Bradford Condon[2], Abdullah Almsaeed[2], Ming Chen[2], Chanaka Mannapperuma[3], Sook Jung[4] and Stephen Ficklin[4]*

[1] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, United States, [2] Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, Knoxville, TN, United States, [3] Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden, [4] Department of Horticulture, Washington State University, Pullman, WA, United States

Despite tremendous advancements in high throughput sequencing, the vast majority of tree genomes, and in particular, forest trees, remain elusive. Although primary databases store genetic resources for just over 2,000 forest tree species, these are largely focused on sequence storage, basic genome assemblies, and functional assignment through existing pipelines. The tree databases reviewed here serve as secondary repositories for community data. They vary in their focal species, the data they curate, and the analytics provided, but they are united in moving toward a goal of centralizing both data access and analysis. They provide frameworks to view and update annotations for complex genomes, interrogate systems level expression profiles, curate data for comparative genomics, and perform real-time analysis with genotype and phenotype data. The organism databases of today are no longer simply catalogs or containers of genetic information. These repositories represent integrated cyberinfrastructure that support cross-site queries and analysis in web-based environments. These resources are striving to integrate across diverse experimental designs, sequence types, and related measures through ontologies, community standards, and web services. Efficient, simple, and robust platforms that enhance the data generated by the research community, contribute to improving forest health and productivity.

Keywords: database, content management system, forest tree, bioinformatics, web services

OPEN ACCESS

Edited by:
Matias Kirst,
University of Florida, United States

Reviewed by:
Matthew Aaron Gitzendanner,
University of Florida, United States
Jason Holliday,
Virginia Tech, United States

****Correspondence:***
Jill L. Wegrzyn
jill.wegrzyn@uconn.edu
Margaret A. Staton
mstaton1@utk.edu
Nathaniel R. Street
nathaniel.street@umu.se
Dorrie Main
dorrie@wsu.edu
Emily Grau
emily.grau@uconn.edu

Specialty section:
This article was submitted to Plant Biotechnology, a section of the journal Frontiers in Plant Science

Received: *16 August 2018*
Accepted: *05 June 2019*
Published: *25 June 2019*

Citation:
Wegrzyn JL, Staton MA, Street NR, Main D, Grau E, Herndon N, Buehler S, Falk T, Zaman S, Ramnath R, Richter P, Sun L, Condon B, Almsaeed A, Chen M, Mannapperuma C, Jung S and Ficklin S (2019) Cyberinfrastructure to Improve Forest Health and Productivity: The Role of Tree Databases in Connecting Genomes, Phenomes, and the Environment. Front. Plant Sci. 10:813. doi: 10.3389/fpls.2019.00813

INTRODUCTION

Starting in the Sanger sequencing era, significant investments were made to catalog genetic resources in primary repositories (Frishman et al., 1998). EMBL (the European Molecular Biology Laboratory), DDBJ (the DNA Data Bank of Japan), and NCBI (the National Center for Biotechnology Information) GenBank were initiated between 1980 and 1992, and remain freely

accessible and federally funded (Benson et al., 1997; Tateno and Gojobori, 1997). The vast majority of data for these large, sequence-centric databases is sourced directly from researcher submissions that are encouraged through peer review journals. These primary resources have evolved with the data collection and curation needs of today, expanding in terms of both the sequence source and the associated metadata (Sayers et al., 2019). All three specialize in generating persistent identifiers to track a single sequence over an extensive network of resources. A genic identifier, as an example, may link a reference genome in NCBI's Genome, an expression value in the Gene Expression Omnibus (GEO), and support for a UniRef90 cluster. These uniquely accessioned resources are increasingly integrated into secondary and tertiary repositories that subset or enhance these accessions with data specific to the communities they serve (Herrero et al., 2016).

As the data types and experimental designs contributing to these repositories diversified, a plethora of model organism databases (MODs) or clade organism databases (CODs) emerged. These databases sought to provide unique resources for the research communities they serve, through layered curation and specialized integration. The AAtDB (An *Arabidopsis thaliana* Database), developed in 1991 to support the first model plant system, has since evolved into the widely accessed, Arabidopsis Information Resource (TAIR) (Flanders et al., 1998; Huala et al., 2001). Around the same time, USDA-ARS funds were dedicated to developing some of the first informatic portals for economically important crop species, including RiceGenes (Cartinhour, 1997), GrainGenes (Triticeae) (Carollo et al., 2005), MaizeGDB (Lawrence et al., 2004), SoyBase (Grant et al., 2009), and the Dendrome Project for forest trees (Wegrzyn et al., 2008). Some of these databases remain independent funded entities, while others have merged into larger repositories or broadened their scope. There are hundreds of plant-focused organismal databases acting as secondary repositories today (Lai et al., 2012; Chen et al., 2018). The vast majority have moved beyond genetics and genomics data, providing advanced integration through stock centers, phenotypic evaluation, breeding resources, and metabolomic pathway integration.

Forest trees are unique among species represented in crop databases. The vast majority are long-lived, outcrossing, with extensive natural distributions represented by large, diverse and locally adapted populations (Holliday et al., 2017). They represent species of economic importance and are used for paper, pulp, biofuels, food, and timber production. At the same time, they serve as a foundation for watersheds, biodiversity, and contribute substantially to carbon sequestration with forests covering roughly 30% of the earth's surface (Houghton, 2005). Like many plants, forest trees have complex genomes with challenges associated with ploidy and repetitive content. Additionally, gymnosperm tree genomes are exceptionally large, ranging from 10 to 40 Gbp in size (De La Torre et al., 2014). This, in combination with the need to broadly sample these large and diverse populations, yields limited full genome representation. Of the nearly 60,000 forest tree species, less than 35 are associated with an assembled and annotated reference genome (Neale et al., 2013; Plomion et al., 2016). A view into our primary databases reveal that over 2,000 species are associated with genetic information that is of value to the research community (Sayers et al., 2019).

In the era of high throughput technologies that assess genotype, phenotype, and environmental metrics much faster than we can conceive, the need for well structured, efficient, and well-connected databases is apparent. In a recent survey of over 700 investigators in the biological sciences, access to analytical frameworks, long-term storage, and the ability to integrate data across disparate sources, were of primary concern (Barone et al., 2017). The long-lived nature of trees, their pivotal role in local economies, and role in ecosystem health requires an integrated approach that must leverage datasets from a variety of sources. To meet the needs of the research community, forest tree databases are moving away from independent database structures and toward integrated Content Management Systems (CMS) that support specific, shareable modules for query and analysis (Ficklin et al., 2011; Sundell et al., 2015). There is less focus on standardized database backends since web services allow users to expose their data, and data definitions. The ability to provide this, and cross-site query, relies heavily on the curation of ontologies to describe data housed in these frameworks. Initiatives that curate woody plant ontologies, to describe plant structures and measured traits, are critical components of forest tree database cross-talk (Jaiswal et al., 2005; Lens et al., 2012; Cooper et al., 2018). When these terms are provided within the framework of recommended standards for data collection, such as the Minimal Information About a Plant Phenotyping Experiment (MIAPPE), opportunities for analytical pipelines to evaluate complex data becomes a reality (Krajewski et al., 2015). In addition to these standards, all tree databases are integrated with existing analytical frameworks, such as Galaxy, that support and expose common bioinformatic workflows (Afgan et al., 2018). In this review, four tree databases are described, including their history, scope, current resources, and analytic tools (**Figure 1**).

TREEGENES

TreeGenes[1], previously known as Dendrome, was initially constructed to provide access to genetic data for forest trees in a relational framework. Early development included curation of molecular markers, genetic maps, Expressed Sequence Tags (ESTs), and species range maps. TreeGenes remained in a custom database schema and website, adopting components of the Generic Model Organism (GMOD) framework for housing genetic maps (cMAP) and genome assemblies (JBrowse) (Stein et al., 2002). Later development focused on the integration of genotyping resources, phenotypes, expression studies, and additional reference genome sequences (Wegrzyn et al., 2012; Falk et al., 2018).

TreeGenes currently represents just over 1700 species from 16 orders and 124 genera (Falk et al., 2018). TreeGenes has 1200 registered users with associated colleague accounts that enable access to data submission and analytical pipelines.

[1]http://www.treegenesdb.org

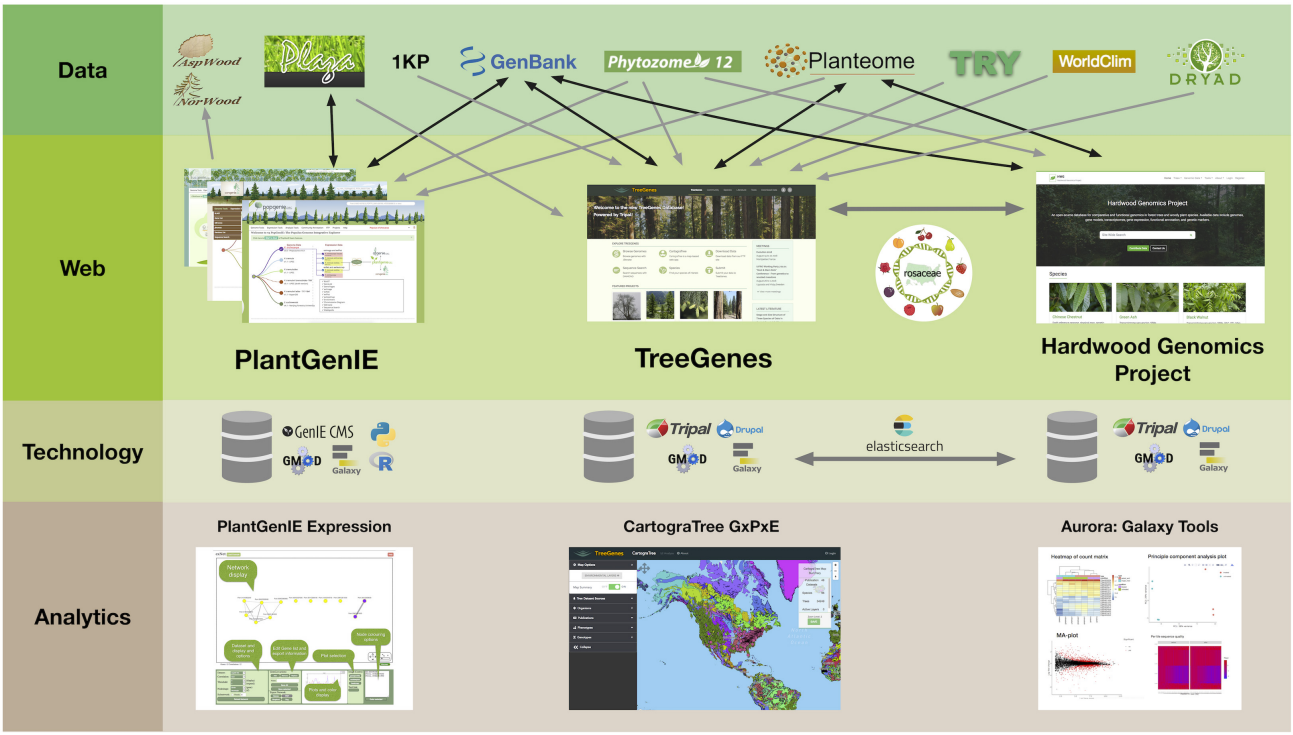

FIGURE 1 | PlantGenIE, TreeGenes, and Hardwood Genomics Project represent integrated web-based frameworks that rely on a combination of primary repositories, secondary plant comparative databases, and user submissions to provide further value through data curation, integration, and analytics.

The database contains 27 reference genomes, 100 genetic maps, 36.7 M genotypes, 303 species with transcriptomes, 40 species with TreeGenes' Unigenes, 306 unique phenotype measures and 935,596 phenotypic measures. Genomic data is sourced primarily from GenBank, 1KP, Phytozome, and PLAZA (Goodstein et al., 2012; Matasci et al., 2014; Proost et al., 2015; Sayers et al., 2019). Phenotypic data is integrated from TRY-DB and Dryad, but primarily comes from direct user submissions (Kattge et al., 2011). Environmental data is extracted from imported layers, including temperature, precipitation, and solar radiation from WorldClim (Fick and Hijmans, 2017), and a variety of metrics from the Harmonized World Soil Survey Database (FAO et al., 2009).

TreeGenes is running on Tripal v3 which integrates a content management system known as Drupal with the Genetic Model Organism Database's (GMOD) relational schema, known as Chado. This conversation in 2017, aligned TreeGenes for the first time with over 30, primarily plant, databases (Ficklin et al., 2011; Sanderson et al., 2013). Recent focused development in Tripal, led by the tree and legume community, enabled cross-site communication, access to efficient data transfer, and the ability to interface with a local installation of Galaxy (Watts and Feltus, 2017; Wytko et al., 2017). Galaxy is an independent framework that provides an API to abstract command line informatic software, develop workflows, and connect to high performance computing resources. Conversion into Tripal resulted in a complete overhaul of the database, and has enabled the development of several analytical modules that allow researchers to search, filter, and funnel data directly into supported workflows.

Following conversion, TreeGenes released a set of Tripal modules that can be utilized by researchers visiting the site or installed and customized for any Tripal supported databases. Tripal Sequence Similarity Search (TSeq) provides access to genomes, transcriptomes, and curated TreeGenes unigenes through traditional NCBI BLAST or optimized Diamond protein searches (Boratyn et al., 2013; Buchfink et al., 2015). The Tripal Plant PopGen Submit (TPPS) module presents a framework for researchers to submit their association genetics, landscape genomics, and related population studies by collecting any combination of molecular markers, phenotypes, and environmental measures. This module implements MIAPPE standards and the associated ontologies to ensure data integrity. The Tripal OrthoQuery module provides a framework for curating unigenes, executing OrthoFinder (Emms and Kelly, 2015), and generating interactive visualizations of gene families in a phylogenetic context. OrthoQuery enables both real-time orthologous gene family analysis and functional assessment of the resulting orthogroups.

Current development in TreeGenes focuses on CartograTree, which enables integration of genotype, phenotype and environmental data for georeferenced forest trees (Herndon et al., 2018). This module provides a robust framework

to query publication datasets, species, phenotypes, genotypes, and associations based on metadata collected at the time of submission. The data and metadata exposed in CartograTree is derived from published population level studies submitted to TreeGenes via TPPS or curated from Dryad. Landscape genomics, association genetics, and population structure analysis is executed through the Galaxy framework.

HARDWOOD GENOMICS PROJECT

The Hardwood Genomics Project (HWG) provides access to genomic resources generated from angiosperm trees, including forest and urban trees of ecological and agricultural significance. The resource originated from the Fagaceae Genomics Web, built in 2007, to house transcriptomes, genomes and genetic maps. As new collaborators joined the effort and the scope of species extended beyond Fagaceae, the site was rebuilt in 2011 as the HWG. HWG's mission is to host reference genomes and transcriptomes that are either not accessible elsewhere, or only available as raw files without an associated and searchable, functional annotation. In addition, HWG accepts molecular markers, genetic maps, germplasm and population descriptions, and community project descriptions. Current resources support species associated with pest or pathogen threats, including green ash, European ash, American chestnut, American beech, black walnut, and redbay, as well as trees with significant economic value, including white oak, black cherry, sugar maple, and tulip poplar.

For species with an available reference genome, HWG provides a workspace for accessing the annotation. This provides value to these sequence resources by performing and hosting functional annotation, including: identification of Open Reading Frames (ORFs) from transcripts, BLAST annotations derived from the Uniprot Swiss-Prot/TrEMBL plant protein databases, InterProScan domain searches (Jones et al., 2014), Gene Ontology (GO) term assignments (Ashburner et al., 2000), and predictions for Simple Sequence Repeats (SSRs) and primers. Researchers can download flat files, explore the spatial context of the assembly with JBrowse (Buels et al., 2016), search functional annotation for genes, and explore assigned GO terms through the ontology graphs. Additional genome specific data, such as alternative splicing, variants, and molecular markers are added to the JBrowse viewer when available.

Hardwood Genomics Project is currently running Tripal v3, and like TreeGenes, is responsible for the development of custom modules that can be installed on other Tripal-enabled sites. RNASeq data is poorly integrated in plant community databases despite the widespread use of expression studies to examine responses to biotic and abiotic stressors in plant systems. To address this limitation, HWG launched a framework devoted to the integration and analysis of gene expression experiments. BioSamples imported from GenBank, with the metadata describing the tissues, treatments, experimental design, and informatic methods, can be explored and compared. Each transcript, examined as part of an RNASeq experiment, has expression values that can be interrogated through interactive visualizations or downloaded for further analysis. The expression data displays can be customized interactively, grouping BioSamples by their tagged metadata values. A tool for comparing gene expression is also available, allowing the user to provide their own gene list and generate a heat map comparing expression of those genes across the relevant BioSamples (Chen et al., 2017). Current development in HWG is focused on supporting bioinformatic workflows, through Galaxy, to allow users to load their own datasets for analysis. HWG has also developed an Elastic Search module that enables search engine style cross-site query. This enables the discovery of relevant datasets within and across Tripal-enabled websites. The Aurora Galaxy Tripal module allows informatic tools to be wrapped in R Markdown which makes it possible to generate Galaxy workflow outputs as static websites.

GENOME DATABASE FOR ROSACEAE

The Genome Database for Rosaceae (GDR[2], Jung et al., 2013) was initiated in 2003 to provide curated and integrated, genomic, genetics and breeding (GGB) data alongside analysis tools to enable basic, translational and applied research. Rosaceae is an economically, nutritionally and biologically important plant family that includes the majority of tree fruits (apple, apricot, blackberry, cherry, peach, plum), nuts (almond), and ornamentals (pear, crab apple). While not specifically focused on forest trees, GDR is included here for its role in developing Tripal modules for breeding and the comparative genomics utility with forest hardwoods.

GDR contains 21 genome assemblies and annotations for 14 species. A total of 528,890 genes, reference transcriptomes for all major species, 14,411 germplasm records, 313 genetic maps, 3.3 M molecular markers, 402,559 phenotype measurements, 3,902 QTL/MTL for 392 agronomic traits, 10.8 M genotypes, and 7,449 publications are housed in the database. GDR provides access to breeding management and analysis tools, pathway analysis through PlantCyc and Pathway Inspector, flexible front-end querying, genome annotations through JBrowse, and sequence similarity search functionality (Jung et al., 2016, 2017). GDR is participating in the development of new Tripal modules; visualization and analysis of genetic maps is available through the new Tripal Map Viewer module while whole genome alignments can be executed through the Tripal Synteny Viewer (Jung et al., 2018). GDR is currently expanding the analytic capabilities of their Breeding Information Management System (BIMS) and developing reference genome integration for the Tripal Map Viewer module.

PLANT GENOME INTEGRATIVE EXPLORER

The Plant Genome Integrative Explorer (PlantGenIE) began as The Populus Integrative Genome Explorer (PopGenIE;

[2]https://www.rosaceae.org

Sjödin et al., 2009), to overcome a lack of tools for routine tasks such as annotating gene lists, converting among sequence identifiers, and visualizing transcript abundance on the basis of EST sequencing. The Populus version was expanded to include visualization of poplar microarray data using the concept of the *Arabidopsis* electronic fluorescent pictograph (eFP) browser (Winter et al., 2007), gene set enrichment tests for Gene Ontology (Ashburner et al., 2000), Pfam (Finn et al., 2010), genome synteny browsing alongside sequence similarity searching. Later, a complimentary comparative co-expression tool was developed to facilitate inference of functional orthologs on the basis of conserved co-expression (Netotea et al., 2014). The resulting networks were integrated within *Populus* and *Arabidopsis* GenIE sites. With the release of the Norway spruce (*Picea abies*) genome (Nystedt et al., 2013), the associated resources were made available in a Conifer database, ConGenIE, which also includes genomes for loblolly pine (*Pinus taeda*; Neale et al., 2014; Wegrzyn et al., 2014) and white spruce (*Picea glauca*; Birol et al., 2013).

The PlantGenIE umbrella resource (Sundell et al., 2015), which included the development of new and updated gene expression tools, together with an integrated gene family analysis, is now available for all species. The primary aim is visualization of gene expression data, primarily from forest tree species, but including related sites for plant models, such as *Arabidopsis*. As such, gene expression resources for aspen (AspWood; Sundell et al., 2017) and Norway spruce (NorWood; Jokipii-Lukkari et al., 2017) cryogenic tangential cuttings series profiling wood development are being integrated within the PopGenIE and ConGenIE sites, respectively. Dedicated sites have been developed to provide access to spatial transcriptomics (Giacomello et al., 2017) and laser capture microdissection (Canas et al., 2014) gene expression data. For a subset of species, community annotation is also provided via WebApollo (Stevens et al., 2016).

The GenIE sites were originally developed using the Drupal CMS with many of the tools from GMOD (Stein et al., 2002). More recently, an alternative open-source CMS has been developed (GenIE-CMS[3]) and the PlantGenIE resource is currently being updated in this platform. GenIE-CMS includes web services, enabling end users to access genomic information from external interfaces such as R and Python analysis scripts. Alongside this update, new and improved versions of gene expression visualization tools have been developed and made available as plugins to GenIE-CMS. The PlantGenIE update includes integration with the PLAZA resource (Proost et al., 2015), integration of cross-GenIE gene lists using PLAZA gene orthology inference methods, new integrative gene expression explorer tools, new and updated gene expression networks inferred using seidr (Schiffthaler et al., 2018), and an updated functional enrichment tool. In addition to updating existing reference genomes, new genomes are being added, including a dedicated eucalyptus site, EucGenIE. The development of GenIE-CMS enables rapid and easy implementation of new GenIE resources and cross-linking among existing GenIE sites using PLAZA gene family and orthology information.

FUTURE DIRECTIONS

Tree database cyberinfrastructure is supporting comparative genomics, population genetics, expression profiling, and genome annotation. These resources focus on a combination of model and non-model systems and integrate with established comparative resources to deliver value added information. Despite their importance, the sustainability of cyberinfrastructure and the related activities of curating and importing scientific data is always in question. The databases described here are leveraging larger open-source projects as their base framework and sharing web-based applications for common functionality, such as genome browsing and sequence similarity searching. For the forest tree community, the majority of the functionality described here has been deployed within the last 3 years and represents the first coordinated effort across these resources. Frameworks like Tripal and PlantGenIE focus on efficient deployment, web services for cross-talk, data visualization, and analytics to provide a robust environment for end users. As an example, the Elastic Search module developed by HWG allows one to search a gene, genome, marker, and other indexed objects in one database and locate results in other Tripal databases without executing independent searches on each website. Sharing development across a larger community allows forest tree databases to focus on the specific needs of their users. Their independent value exists in the additional curation, metadata acquisition, indexing, analytics, and visualization that is not delivered from the primary repositories. TreeGenes and Hardwood Genomics Web, are focused on expression data integration for non-model trees and metadata retrieval and cross-study analytics for population genetics studies. GDR is focused on improving access and visualization of genetic maps as well as breeding tools. The PlantGenIE framework is providing a robust platform for species with a reference genome, and advanced visualization for expression data that integrates across studies. All of these databases are also seeking stronger connections to more broadly plant focused repositories, such as Phytozome, PLAZA, and Planteome, that provide genetic and ontological resources that improve the utility of cross-site querying.

While tremendous advancements have been made through recent and focused development on these pivotal frameworks, several challenges remain for the forest tree community. As datasets become larger and more integrative, it is increasingly difficult for small database teams to keep up with the data capture and curation. With increasing access to reference genomes, large-scale population studies, and high throughput environmental data, biological databases must develop more efficient metadata capture, storage, and query capacity. These repositories will be tasked with implementing advanced natural language processing, automated metadata capture, and ontological term assignment to span not only genetic data, but associated phenotypic and environmental data. These latter categories encompass an expansive range, from traditional growth traits to canopy

[3]https://github.com/PlantGenIE/GenIECMS

metrics, soil profiles, microbiomes, and metatranscriptomics. The biomedical community has paved the way for some of this technology but forest tree data, and the associated genetic resources, remain more heterogenous (Koleck et al., 2019). This heterogeneity is combined with high throughput technologies, such as remote sensing, that challenge existing cyberinfrastructure in terms of efficient transfer, storage, and query (Côté et al., 2018). Capturing data for large forest tree populations may involve storing millions of genotypes across thousands of individuals or hundreds of pangenomes. It will also rely on a combination of sequencing and phenotyping technologies that continue to evolve (Bolger et al., 2019). After the storage and minimal reporting requirements are established, the frameworks the databases are built upon will need to assist users in determining the most appropriate analytics and provide the required formatting for the queried data. While progress has been made in connecting data to workflows on high performance computing, such as Galaxy; systems that can recommend appropriate workflows are still in progress. The future of biological databases for all plants is reproducible workflows that represent the metadata associated with the original studies. Concerted efforts in this area and integration of new data types evolving from high throughput technologies will be key to advancing discovery for the forest tree community.

AUTHOR CONTRIBUTIONS

JW, MS, NS, DM, and SF designed the databases and software described. EG, NH, SB, TF, SZ, RR, PR, and LS developed the core TreeGenes and TreeGenes Tripal modules. MS, BC, AA, and MC developed the core HWG and HWG Tripal modules. NS and CM developed the core PlantGenIE. SJ developed the core GDR and GDR Tripal modules. SF developed the core Tripal. JW, MS, NS, and DM wrote the manuscript. All authors read and approved the manuscript.

FUNDING

We would like to acknowledge the funding provided through the National Science Foundation (ACI-1443040 and ACI-1444573) and United States Department of Agriculture (2016-67013-24469).

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., et al. (2018). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi: 10.1093/nar/gky379

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.* 25, 25–29.

Barone, L., Williams, J., and Micklos, D. (2017). Unmet needs for analyzing biological big data: a survey of 704 NSF principal investigators. *PLoS Comp. Biol.* 13:e1005755. doi: 10.1371/journal.pcbi.1005755

Benson, D. A., Boguski, M. S., Lipman, D. J., and Ostell, J. (1997). GenBank. *Nucleic Acids Res.* 25, 1–6. doi: 10.1093/nar/25.1.1

Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., et al. (2013). Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics* 29, 1492–1497. doi: 10.1093/bioinformatics/btt178

Bolger, A. M., Poorter, H., Dumschott, K., Bolger, M. E., Arend, D., Osorio, S., et al. (2019). Computational aspects underlying genome to phenome analysis in plants. *Plant J.* 97, 182–198. doi: 10.1111/tpj.14179

Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., et al. (2013). BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* 41, W29–W33. doi: 10.1093/nar/gkt282

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nat. Methods* 12:59. doi: 10.1038/nmeth.3176

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17:66. doi: 10.1186/s13059-016-0924-1

Canas, R. A., Canales, J., Gomez-Maldonado, J., Avila, C., and Canovas, F. M. (2014). Transcriptome analysis in maritime pine using laser capture microdissection and 454 pyrosequencing. *Tree Physiol.* 34, 1278–1288. doi: 10.1093/treephys/tpt113

Carollo, V., Matthews, D. E., Lazo, G. R., Blake, T. K., Hummel, D. D., Lui, N., et al. (2005). GrainGenes 2.0. an improved resource for the small-grains community. *Plant Physiol.* 139, 643–651. doi: 10.1104/pp.105.064485

Cartinhour, S. W. (1997). Public informatics resources for rice and other grasses. *Plant Mol. Biol.* 35, 241–251. doi: 10.1007/978-94-011-5794-0_23

Chen, F., Dong, W., Zhang, J., Guo, X., Chen, J., Wang, Z., et al. (2018). The sequenced angiosperm genomes and genome databases. *Front. Plant Sci.* 9:418. doi: 10.3389/fpls.2018.00418

Chen, M., Henry, N., Almsaeed, A., Zhou, X., Wegrzyn, J., Ficklin, S., et al. (2017). New extension software modules to enhance searching and display of transcriptome data in tripal databases. *Database* 2017:bax052. doi: 10.1093/database/bax052

Cooper, L., Meier, A., Laporte, M.-A., Elser, J. L., Mungall, C., Sinn, B. T., et al. (2018). The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res.* 46, D1168–D1180. doi: 10.1093/nar/gkx1152

Côté, J. F., Fournier, R. A., Luther, J. E., and van Lier, O. R. (2018). Fine-scale three-dimensional modeling of boreal forest plots to improve forest characterization with remote sensing. *Remote Sens. Environ.* 219, 99–114. doi: 10.1016/j.rse.2018.09.026

De La Torre, A., Birol, I., Bousquet, J., Ingvarsson, P., Jansson, S., Jones, S. J. M., et al. (2014). Insights into conifer giga-genomes. *Plant Physiol.* 166, 1724–1732. doi: 10.1104/pp.114.248708

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Falk, T., Herndon, N., Grau, E., Buehler, S., Richter, P., Zaman, S., et al. (2018). Growing and cultivating the forest genomics database, TreeGenes. *Database* 2018:bay084. doi: 10.1093/database/bay084

FAO, IIASA and ISSCAS, ISRIC (2009). *JRC Harmonized World Soil Database (Version 1.1)*. Rome: FAO.

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1038/sdata.2018.254

Ficklin, S. P., Sanderson, L.-A., Cheng, C.-H., Staton, M. E., Lee, T., Cho, I.-H., et al. (2011). Tripal: a construction toolkit for online genome databases. *Database* 2011:bar044. doi: 10.1093/database/bar044

Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., et al. (2010). The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222. doi: 10.1093/nar/gkp985

Flanders, D. J., Weng, S., Petel, F. X., and Cherry, J. M. (1998). AtDB, the *Arabidopsis thaliana* database, and graphical-web-display of progress by the

Arabidopsis genome initiative. *Nucleic Acids Res.* 26, 80–84. doi: 10.1093/nar/26.1.80

Frishman, D., Heumann, K., Lesk, A., and Mewes, H. W. (1998). Comprehensive, comprehensible, distributed and intelligent databases: current status. *Bioinformatics* 14, 551–561. doi: 10.1093/bioinformatics/14.7.551

Giacomello, S., Salmén, F., Terebieniec, B. K., Vickovic, S., Navarro, J. F., Alexeyenko, A., et al. (2017). Spatially resolved transcriptome profiling in model plant species. *Nat. Plants* 3:17061. doi: 10.1038/nplants.2017.61

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2009). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

Herndon, N., Falk, T., Grau, E. S., Jung, S., Ficklin, S., Main, D., et al. (2018). "Association mapping for forest trees with CartograTree," in *Application of Semantic Technologies in Biodiversity Science. Studies on the Semantic Web*, ed. A. E. Thessen (Amsterdam: IOS Press).

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., et al. (2016). Ensembl comparative genomics resources. *Database* 2016:bav096. doi: 10.1093/database/bav096

Holliday, J. A., Aitken, S. N., Cooke, J. E. K., Fady, B., Gonzalez-Martinez, S. C., Heuertz, M., et al. (2017). Advances in ecological genomics in forest trees and applications to genetic resources conservation and breeding. *Mol. Ecol.* 26, 706–717. doi: 10.1111/mec.13963

Houghton, R. A. (2005). Aboveground forest biomass and the global carbon balance. *Glob. Change Biol.* 11, 945–958. doi:10.1111/j.1365-2486.2005.00955.x

Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., et al. (2001). The *Arabidopsis* information resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 29, 102–105. doi: 10.1093/nar/29.1.102

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., et al. (2005). Plant ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics* 6, 388–397. doi: 10.1002/cfg.496

Jokipii-Lukkari, S., Sundell, D., Nilsson, O., Hvidsten, T. R., Street, N. R., and Tuominen, H. (2017). NorWood: a gene expression resource for evo-devo studies of conifer wood development. *New Phytol.* 216, 482–494. doi: 10.1111/nph.14458

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jung, S., Ficklin, S. P., Lee, T., Cheng, C.-H., Blenda, A., Zheng, P., et al. (2013). The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res.* 42, D1237–D1244. doi: 10.1093/nar/gkt1012

Jung, S., Lee, T., Cheng, C. H., Buble, K., Zheng, P., Yu, J., et al. (2018). 15 years of GDR: new data and functionality in the genome database for rosaceae. *Nucleic Acids Res.* 47, D1137–D1145. doi: 10.1093/nar/gky1000

Jung, S., Lee, T., Cheng, C.-H., Ficklin, S., Yu, J., Humann, J., et al. (2017). Extension modules for storage, visualization and querying of genomic, genetic and breeding data in tripal databases. *Database* 2017:bax092. doi: 10.1093/database/bax092

Jung, S., Lee, T., Ficklin, S., Yu, J., Cheng, C.-H., and Main, D. (2016). Chado use case: storing genomic, genetic and breeding data of *Rosaceae* and *Gossypium* crops in Chado. *Database* 2016:baw010. doi: 10.1093/database/baw010

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., et al. (2011). TRY – a global database of plant traits. *Glob. Change Biol.* 17, 2905–2935. doi: 10.1111/j.1365-2486.2011.02451.x

Koleck, T. A., Dreisbach, C., Bourne, P. E., and Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J. Am. Med. Inform. Assoc.* 26, 364–379. doi: 10.1093/jamia/ocy173

Krajewski, P., Chen, D., Ćwiek, H., van Dijk, A. D. J., Fiorani, F., Kersey, P., et al. (2015). Towards recommendations for metadata and data handling in plant phenotyping. *J. Exp. Bot.* 66, 5417–5427. doi: 10.1093/jxb/erv271

Lai, K., Lorenc, M. T., and Edwards, D. (2012). Genomic databases for crop improvement. *Agronomy* 2, 62–73. doi: 10.3390/agronomy2010062

Lawrence, C. J., Dong, Q., Polacco, M. L., Seigfried, T. E., and Brendel, V. (2004). MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.* 32, D393–D397. doi: 10.1093/nar/gkh011

Lens, F., Cooper, L., Gandolfo, M. A., Groover, A., Jaiswal, P., Lachenbruch, B., et al. (2012). An extension of the plant ontology project supporting wood anatomy and development research. *IAWA J.* 33, 113–117. doi: 10.1163/22941932-90000083

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience* 3:17. doi: 10.1186/2047-217X-3-17

Neale, D. B., Langley, C. H., Salzberg, S. L., and Wegrzyn, J. L. (2013). Open access to tree genomes: the path to a better forest. *Genome Biol.* 14:120. doi: 10.1186/gb-2013-14-6-120

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., et al. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* 15:R59. doi: 10.1186/gb-2014-15-3-r59

Netotea, S., Sundell, D., Street, N. R., and Hvidsten, T. R. (2014). ComPlEx: conservation and divergence of co-expression networks in *A. thaliana, Populus* and *O. sativa*. *BMC Genomics* 15:106. doi: 10.1186/1471-2164-15-106

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584.

Plomion, C., Bastien, C., Bogeat-Triboulot, M.-B., Bouffier, L., Déjardin, A., Duplessis, S., et al. (2016). Forest tree genomics: 10 achievements from the past 10 years and future prospects. *Ann. For. Sci.* 73, 77–103. doi: 10.1007/s13595-015-0488-3

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., et al. (2015). PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.* 43, D974–D981. doi: 10.1093/nar/gku986

Sanderson, L.-A., Ficklin, S. P., Cheng, C.-H., Jung, S., Feltus, F. A., Bett, K. E., et al. (2013). Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database* 2013:bat075. doi: 10.1093/database/bat075

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., and Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Res.* 47, D94–D99. doi: 10.1093/nar/gky989

Schiffthaler, B., Serrano, A., Delhomme, N., and Street, N. R. (2018). Seidr: a toolkit for calculation of crowd networks. *bioRxiv*

Sjödin, A., Street, N. R., Sandberg, G., Gustafsson, P., and Jansson, S. (2009). The populus genome integrative explorer (PopGenIE): a new resource for exploring the populus genome. *New Phytol.* 182, 1013–1025. doi: 10.1111/j.1469-8137.2009.02807.x

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., et al. (2002). The generic genome browser: a building block for a model organism system database. *Genome Res.* 12, 1599–1610. doi: 10.1101/gr.403602

Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., et al. (2016). Sequence of the sugar pine megagenome. *Genetics* 204, 1613–1626. doi: 10.1534/genetics.116.193227

Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.-C., Sjödin, A., et al. (2015). The plant genome integrative explorer resource: plantgenIE.org. *New Phytol.* 208, 1149–1156. doi: 10.1111/nph.13557

Sundell, D., Street, N. R., Kumar, M., Mellerowicz, E. J., Kucukoglu, M., Johnsson, C., et al. (2017). AspWood: high-spatial-resolution transcriptome profiles reveal uncharacterized modularity of wood

formation in *Populus tremula*. *Plant Cell* 29, 1585–1604. doi: 10.1105/tpc.17.00153

Tateno, Y., and Gojobori, T. (1997). DNA data bank of Japan in the age of information biology. *Nucleic Acids Res.* 25, 14–17. doi: 10.1093/nar/25.1.14

Watts, N. A., and Feltus, F. A. (2017). Big Data Smart Socket (BDSS): a system that abstracts data transfer habits from end users. *Bioinformatics* 33, 627–628. doi: 10.1093/bioinformatics/btw679

Wegrzyn, J. L., Lee, J. M., Tearse, B. R., and Neale, D. B. (2008). TreeGenes: a forest tree genome database. *Int. J. Plant Genomics* 2008:412875. doi: 10.1155/2008/412875

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L.-S., Loopstra, C. A., Vasquez-Gross, H. A., et al. (2014). Unique features of the loblolly pine (*Pinus taeda L.*) megagenome revealed through sequence annotation. *Genetics* 196, 891–909. doi: 10.1534/genetics.113.159996

Wegrzyn, J. L., Main, D., Figueroa, B., Choi, M., Yu, J., Neale, D. B., et al. (2012). Uniform standards for genome databases in forest and fruit trees. *Tree Genet. Genomes* 8:549. doi: 10.1007/s11295-012-0494-7

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V., and Provart, N. J. (2007). An 'electronic fluorescent pictograph' browser for exploring and analyzing large-scale biological data sets. *PLoS One* 2:e718. doi: 10.1371/journal.pone.0000718

Wytko, C., Soto, B., and Ficklin, S. P. (2017). blend4php: a PHP API for galaxy. *Database* 2017:baw154. doi: 10.1093/database/baw154

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.