CartograTree: Enabling Landscape Genomics for Forest Trees

Nic Herndon[1], Emily S. Grau[1], Iman Batra[1], Steven A. Demurjian Jr.[1], Hans A. Vasquez-Gross[2], Margaret E. Staton[3], and Jill L. Wegrzyn[1]

[1]University of Connecticut, Department of Ecology and Evolutionary Biology, Storrs, CT, USA
[2]University of California, Davis, Department of Plant Sciences, Davis, CA, USA
[3]University of Tennessee, Department of Entomology and Plant Pathology, Knoxville, TN, USA

ABSTRACT

Forest trees cover just over 30% of the earth's surface and are studied by researchers around the world for both their conservation and economic value. With the onset of high throughput technologies, tremendous phenotypic and genomic data sets have been generated for hundreds of species. These long-lived and immobile individuals serve as ideal models to assess population structure and adaptation to environment. Despite the availability of comprehensive data, researchers are challenged to integrate genotype, phenotype, and environment in one place. Towards this goal, CartograTree was designed and implemented as a repository and analytic framework for genomic, phenotypic, and environmental data for forest trees. One of key components, the integration of geospatial data, allows the display of environmental layers and acquisition of environmental metrics relative to the positions of georeferenced individuals.

Keywords: genotype, phenotype, association mapping, forest trees, genomics, environment, GIS

BACKGROUND

Forest trees have a major ecological and economic impact worldwide. They are critical for CO_2 sequestration, prevention of soil erosion, maintenance of watersheds, and promotion of biodiversity. From an economic point of view, they are necessary for timber production as well as an alternative source of biofuels. In light of climate change, populations of forest trees are being subjected to longer droughts, as well as increased damage from introduced and native pests and pathogens. From a production perspective, there is a desire to improve breeding populations through the identification of loci contributing to the growth and wood quality traits of interest.

Association genetics identifies correlations between genotypes and phenotypes in large populations. In other words, it finds the relationship between a set of genes or genome regions that contribute to a trait or a disease. Next generation sequencing technologies have led to increasing availability of genomic and transcriptomic sequences. They have shifted the bottleneck in omics studies from data generation to data storage and analysis. High-throughput genotyping and phenotyping is now routinely performed on thousands individuals at a time through genome-wide association studies (GWAS). The associations sought in forest trees are further enabled by their large, diverse outcrossing populations as well as their extended longevity in the same environment. This allows the association of genotype to phenotype as well as genotype to environment.

To date, researchers have been hampered by the obstacles presented by these large data sets in addition to an inability to integrate across disciplines. CartograTree (Vasquez-Gross et al., 2013) was developed from a collaboration among forest tree ecologists, geneticists, breeders, and physiologists who identified integration and analysis as the critical bottlenecks. CartograTree aims to integrate genomic and phenotypic data for forest trees, along with environmental data through a map-based visualization platform. The tool does not only enable visualization and search capabilities, but allow the researchers to send data for association mapping analysis through the use of semantic technologies and high performance computing resources offered by Cyverse (Goff et al., 2011).

CARTOGRATREE DATA

Currently, two primary repositories are responsible for the acquisition and curation of vast genomics resources. The TreeGenes database (`http://treegenesdb.org`, Wegrzyn et al. (2008, 2012)) currently hosts 13 genome assemblies, transcriptome resources for 262 species, 95 genetic maps, over 110 million genotypes, and nearly 200,000 phenotypic evaluations. While TreeGenes hosts data for over 1200 species, the most substantial genotype and phenotype resources have been curated for conifer species. The Hardwood Genomics Web (HWG, `http://www.hardwoodgenomics.org`, Sanderson et al. (2013)) houses deep RNASeq data from phylogenetically diverse forest tree species. In addition, low coverage genome sequence data, resulting genotypes, and genomic SSRs are available for key hardwood species. Permanent mapping and reference populations (genetic linkage maps) are in development for green ash, tulip poplar, honeylocust, black walnut and northern red oak.

The genotype and phenotype data is delivered to the databases through the TreeGenes Data Repository (TGDR) which takes the scientist through a vetted workflow that collects both association raw data and experimental metadata on existing studies. The majority of the individual tree accessions are associated with georeference coordinates that are requested during the submission. Data is also made available through collaborative studies whose data storage is facilitated by TreeGenes or HWG. Following submission, public studies or datasets are available for geospatial presentation and access in CartograTree.

In addition to the phenotypes delivered through TGDR, TreeGenes collects phenotypes for species with genotype data that also have geo-referenced individuals in TRY-DB. The TRY initiative is a collection of both published and unpublished datasets from a wide variety of trait databases, including: LEDA, GlopNET, BiolFlor, SID, and EcoFlora (Kattge et al., 2011).

Environmental data can be accessed via the WorldClim dataset which contains summarized temperatures and precipitations, and biologically relevant variables for past, current, and future conditions (Hijmans et al., 2005). In addition, the CartograTree application fully integrates the Ameriflux project which includes a total of 156 stations spread across North and South America (`http://ameriflux.lbl.gov`). This information will be coupled with metrics from user-submissions and collaborative projects which may be tracking specific information, such as soil samples.

A system overview of the CartograTree is shown in Figure 1.

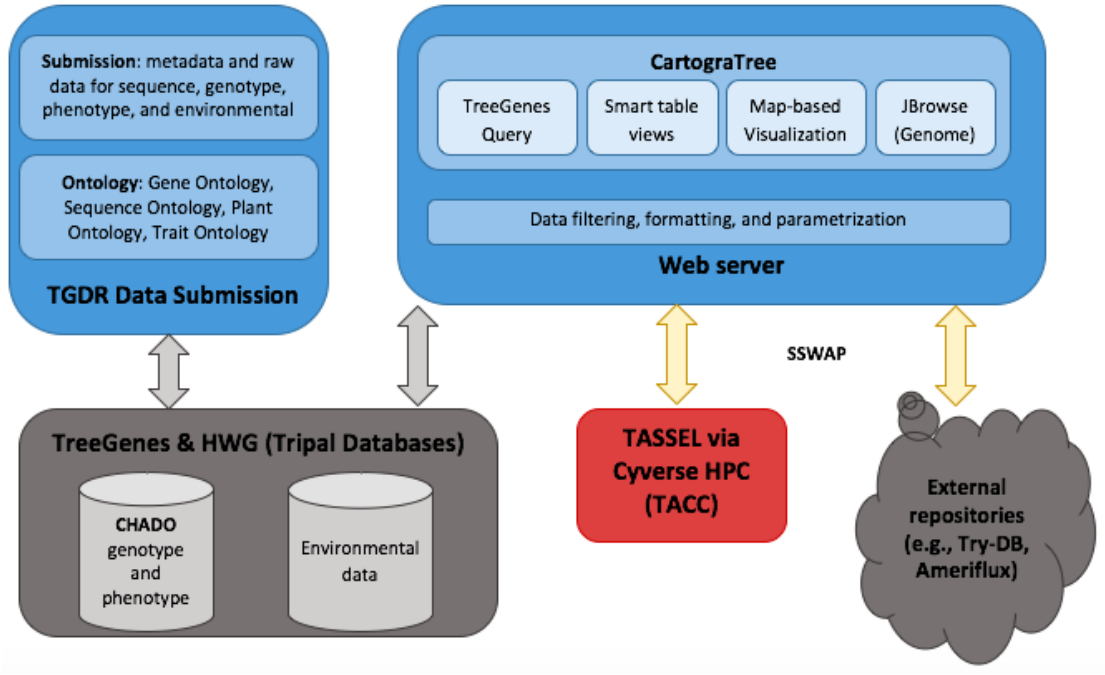

Figure 1. Architectural view of CartograTree. Users can upload data into the Tripal databases using the TreeGenes Data Repository module (TGDR). These databases and other external repositories can be queried and the results visualized using the web interfaces. Users can also further analyze the data using TASSEL.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2345v2 | CC BY 4.0 Open Access | rec: 9 Aug 2016, publ: 9 Aug 2016

CARTOGRATREE SERVICES

CartograTree allows scientists to view georeferenced tree accessions, along with metadata via the Google Maps API, as shown in Figure 2. The web interface allows one to select specific studies, species, regions, specific phenotypes, and more. Researchers can select, filter, combine, and inspect data through basic and more complex queries. Further data exploration is enabled through interaction with web services in the form of the SSWAP semantic discovery pipeline. Data is semantically tagged as a set, and formatted appropriately for analytical applications. Currently, the platform supports multiple sequence alignment, phylogenetic analysis, and association genetics. TASSEL, the application of choice for association genetics, can operate on various combinations of genotypic, phenotypic, and environmental files (Bradbury et al., 2007). Web service entry points are semantically tagged for TASSEL and hosted at the Texas Advanced Computing Center's (TACC) High Performance Computing resources using the Agave API (`http://agaveapi.co`). Semantic TASSEL is automatically discovered by SSWAP (Gessler et al., 2009), and the association data is marshaled to the service. This is a drag-and-drop operation for the scientist negates the need to enter file names, stage, or manually copy the data. The scientist initiates the association mapping analysis, SSWAP monitors the job, and places the final results in the data store hosted by Cyverse.

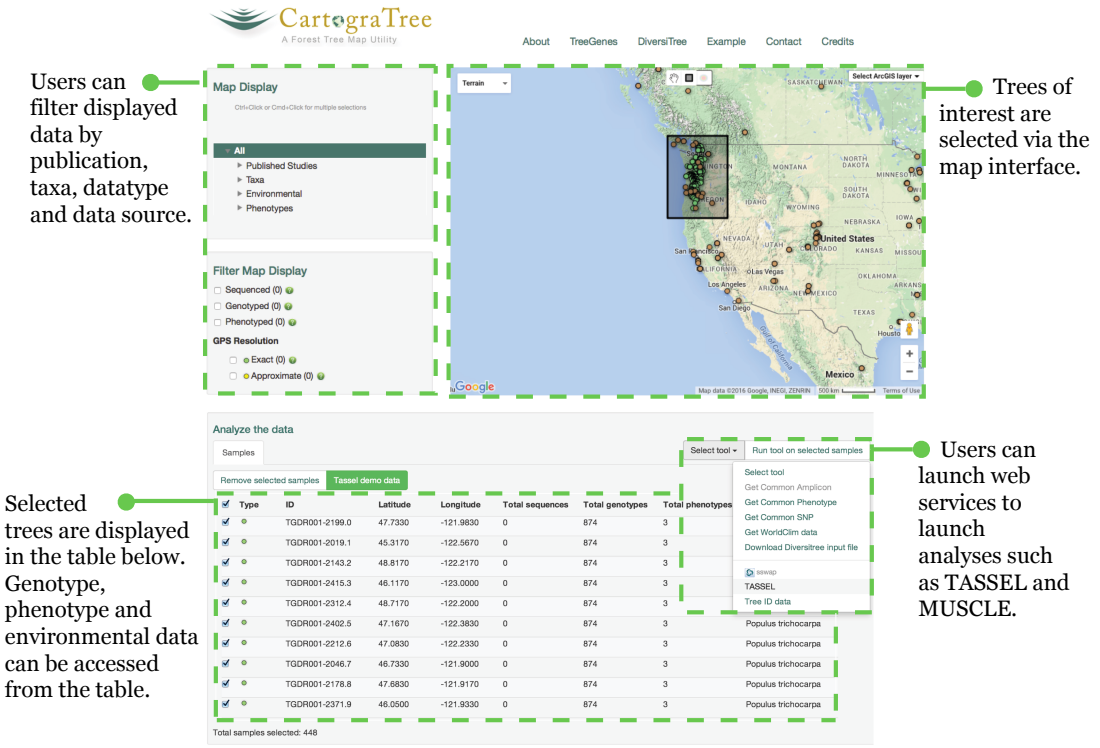

Figure 2. User interface of CartograTree. Users can select the trees of interest using the GUI, filter them through the web form, and select trees for further analyses.

FUTURE DEVELOPMENTS

While CartograTree presents an important framework for data integration and association mapping analysis, further development is needed to ensure its utility to the broader community. Current work is focused on developing a more advanced geospatial framework, including performance, metrics, and the ability to query across high resolution GIS layers. In addition, we are addressing the need for more parameterization of the association mapping runs. This should include the ability to generate population structure information, select a variety of statistical models, and easily filter for missing data or phenotypic distributions. This collaborative project between TreeGenes and Hardwood Genomics Web will enable researchers to integrate valuable and massive datasets to address the numerous threats facing forests worldwide.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2345v2 | CC BY 4.0 Open Access | rec: 9 Aug 2016, publ: 9 Aug 2016

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Award Numbers DBI-0735191 and DBI-1265383 (URL: www.cyverse.org), and Award Number ACI-1443040.

The authors would like to thank Damian Gessler (Semantic Options, LLC) for his continued development and support of SSWAP.

REFERENCES

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635.

Gessler, D. D., Schiltz, G. S., May, G. D., Avraham, S., Town, C. D., Grant, D., and Nelson, R. T. (2009). SSWAP: A simple semantic web architecture and protocol for semantic web services. *BMC bioinformatics*, 10(1):309.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., et al. (2011). The iPlant collaborative: cyberinfrastructure for plant biology. *Frontiers in plant science*, 2:34.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15):1965–1978.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van Bodegom, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., Atkin, O., Bahn, M., Baker, T. R., Baldocchi, D., Bekker, R., Blanco, C. C., Blonder, B., Bond, W. J., Bradstock, R., Bunker, D. E., Casanoves, F., Cavender-Bares, J., Chambers, J. Q., Chapin III, F. S., Chave, J., Coomes, D., Cornwell, W. K., Craine, J. M., Dobrin, B. H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W. F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G. T., Fyllas, N. M., Gallagher, R. V., Green, W. A., Gutierrez, A. G., Hickler, T., Higgins, S. I., Hodgson, J. G., Jalili, A., Jansen, S., Joly, C. A., Kerkhoff, A. J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J. M. H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T. D., Leishman, M., Lens, F., Lenz, T., Lewis, S. L., Lloyd, J., Llusià, J., Louault, F., Ma, S., Mahecha, M. D., Manning, P., Massad, T., Medlyn, B. E., Messier, J., Moles, A. T., Müller, S. C., Nadrowski, K., Naeem, S., Niinemets, U., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V. G., Onoda, Y., Ordoñez, J., Overbeck, G., Ozinga, W. A., Patiño, S., Paula, S., Pausas, J. G., Peñuelas, J., Phillips, O. L., Pillar, V., Poorter, H., Poorter, L., Poschlod, P., Prinzing, A., Proulx, R., Rammig, A., Reinsch, S., Reu, B., Sack, L., Salgado-Negret, B., Sardans, J., Shiodera, S., Shipley, B., Siefert, A., Sosinski, E., Soussana, J.-F., Swaine, E., Swenson, N., Thompson, K., Thornton, P., Waldram, M., Weiher, E., White, M., White, S., Wright, S. J., Yguel, B., Zaehle, S., Zanne, A. E., and Wirth, C. (2011). Try – a global database of plant traits. *Global Change Biology*, 17(9):2905–2935.

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

Vasquez-Gross, H. A., Yu, J. J., Figueroa, B., Gessler, D. D., Neale, D. B., and Wegrzyn, J. L. (2013). CartograTree: connecting tree genomes, phenotypes and environment. *Molecular ecology resources*, 13(3):528–537.

Wegrzyn, J., Main, D., Figueroa, B., Choi, M., Yu, J., Neale, D., Jung, S., Lee, T., Stanton, M., Zheng, P., et al. (2012). Uniform standards for genome databases in forest and fruit trees. *Tree genetics & genomes*, 8(3):549–557.

Wegrzyn, J. L., Lee, J. M., Tearse, B. R., and Neale, D. B. (2008). TreeGenes: a forest tree genome database. *International journal of plant genomics*, 2008.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2345v2 | CC BY 4.0 Open Access | rec: 9 Aug 2016, publ: 9 Aug 2016