TreeGenes
&
Tripal

treegenesdb.org

Emily Grau
Department of Ecology & Evolutionary Biology
University of Connecticut, Storrs CT

TreeGenes Database: History

treegenesdb.org

- Began as the Dendrome project (USDA funded initiative) in 1993 to hold forest tree genetic maps and associated markers
- One of the 1st UDSA funded databases on the internet
- Schema has been changing and evolving ever since

TreeGenes
A Forest Tree Genome Database

TreeGenes Database: History

treegenesdb.org

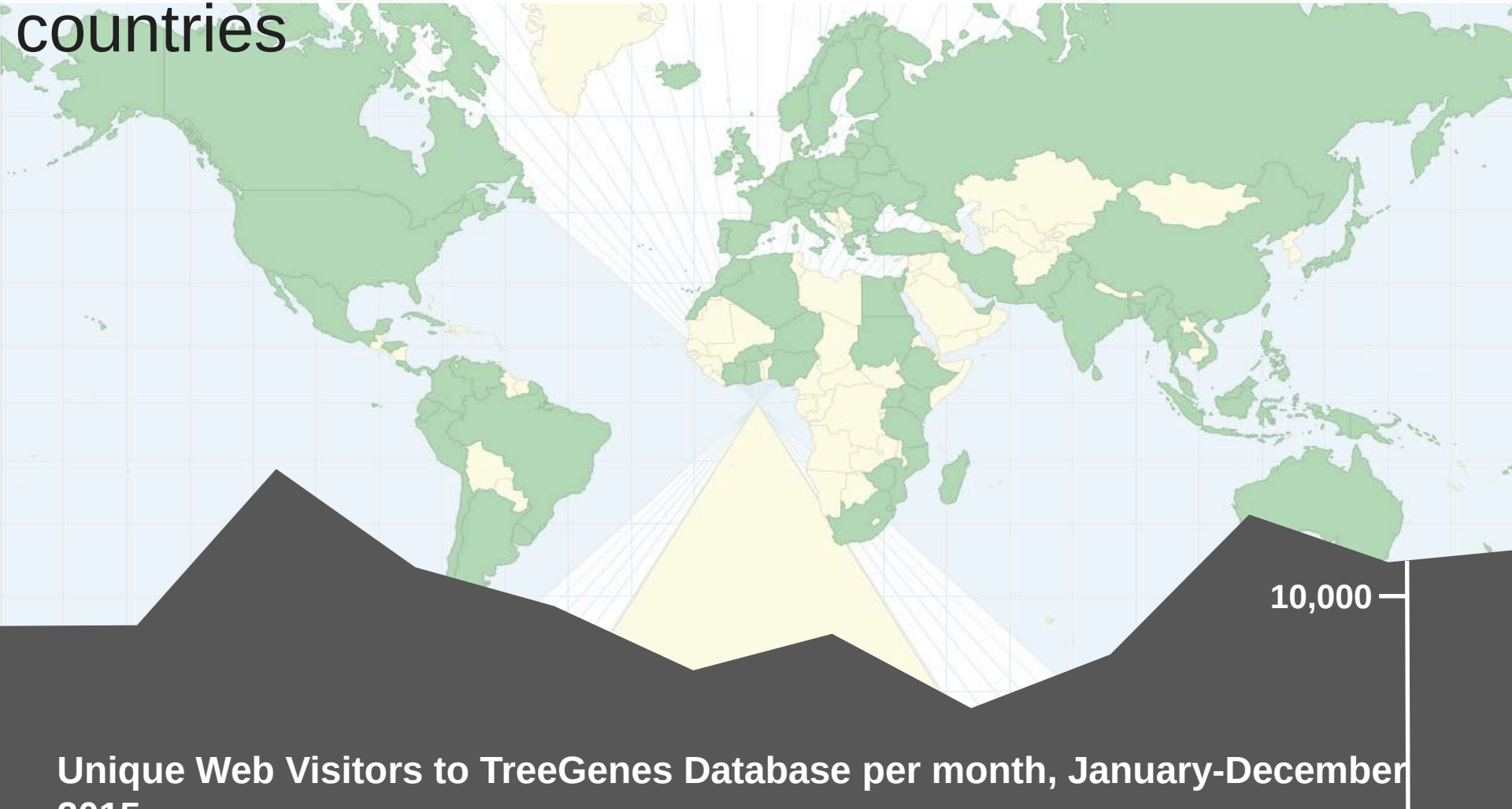
- Began to hold forest tree genetic maps and associated markers
- Expanded to other data types
  - Sequence
    - Reseqeuncing, Large-Scale Genotyping, Transcriptomics/Expression
    - Full Genome Sequences
  - Analysis and Visualization Tools
    - Ability for users to mine the data
  - Resources for the user community
    - Literature, Colleagues

TreeGenes
A Forest Tree Genome Database

TreeGenes Database: Users

treegenesdb.org

2,086 users from 862 organizations in 94 countries



Unique Web Visitors to TreeGenes Database per month, January-December 2015

TreeGenes Database: Species

treegenesdb.org

- 1,774 species from 101 genera
  - At least one genetic artifact from each species
  - Conifers but is also inclusive of all forest trees
- Full genome sequence: 13 species
- Transcriptome/Expression resources: 3,920,817 sequences from 263 species
- 106 genetic maps from 35 species

TreeGenes Database: Data Sources

Automated scripts

- Primary databases such as NCBI
- Appropriate data should be submitted to primary databases **first**

User submissions

- For data and metadata not captured well by primary databases

TreeGenes Database: Data Sources

Automated scripts

- NCBI
  - Transcripts, Protein, Unigene Databases
  - Linked to literature records, etc.
  - Incorporated into visualization tools
- Literature
  - Web of Science, PubMed

```
– <PubmedArticleSet>
  – <PubmedArticle>
    – <MedlineCitation Owner="NLM" Status="PubMed-not-MEDLINE">
      <PMID Version="1">25202539</PMID>
    – <DateCreated>
        <Year>2014</Year>
        <Month>09</Month>
        <Day>09</Day>
      </DateCreated>
    – <DateCompleted>
        <Year>2014</Year>
        <Month>09</Month>
        <Day>09</Day>
      </DateCompleted>
    – <DateRevised>
        <Year>2014</Year>
        <Month>09</Month>
        <Day>12</Day>
      </DateRevised>
    – <Article PubModel="Electronic-eCollection">
      – <Journal>
          <ISSN IssnType="Electronic">2168-0450</ISSN>
        – <JournalIssue CitedMedium="Print">
            <Volume>1</Volume>
            <Issue>4</Issue>
          – <PubDate>
              <Year>2013</Year>
              <Month>Apr</Month>
            </PubDate>
          </JournalIssue>
          <Title>Applications in plant sciences</Title>
          <ISOAbbreviation>Appl Plant Sci</ISOAbbreviation>
        </Journal>
      – <ArticleTitle>
          Application of proteomics to the study of pollination drops
        </ArticleTitle>
      – <Pagination>
          <MedlinePgn/>
        </Pagination>
        <ELocationID EIdType="doi" ValidYN="Y">10.3732/apps
        <ELocationID EIdType="pii" ValidYN="Y">apps.130008
      – <Abstract>
          <AbstractText Label="UNLABELLED">•</AbstractTe
        – <AbstractText Label="PREMISE OF THE STUDY" Nlm
            Pollination drops are a formative component in gymnosp
            sexual reproduction. •
          </AbstractText>
        – <AbstractText Label="METHODS" NlmCategory="ME
            Pollination drops were sampled from eight gymnosperm
```

TreeGenes Database: Data Sources

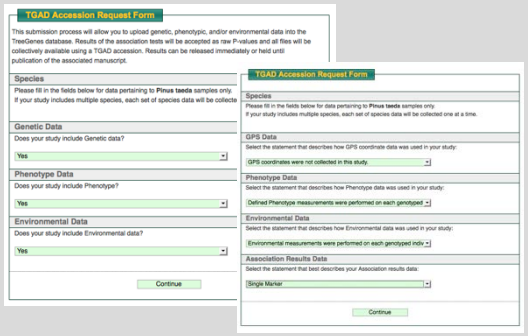
User submissions

- Internal projects or collaborations (day one)
- Submissions of data post-analysis at publication time

TreeGenes Database: Data Sources

treegenesdb.org

User submissions



Submit genetic maps, association or population study data

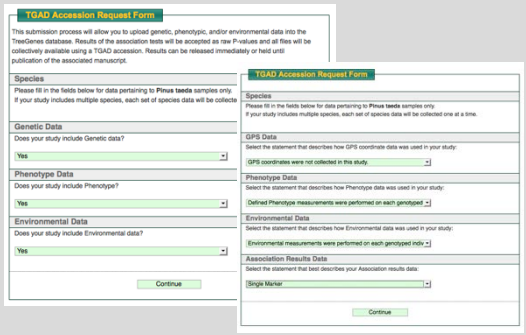
Most submissions from journal requirement: Tree Genetics and Genomes

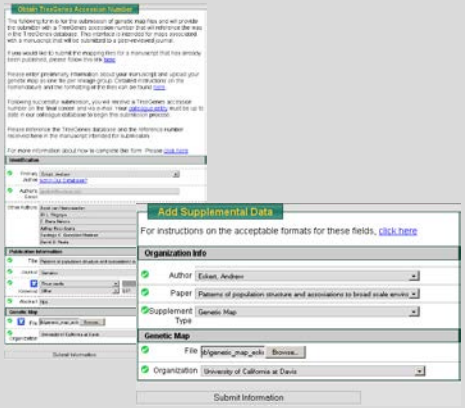
TreeGenes Database: Data Sources

treegenesdb.org

User submissions

Genetic maps, association or population studies





TreeGenes Data Repository

A listing of data submissions is displayed below.
To submit data to TreeGenes, click here.

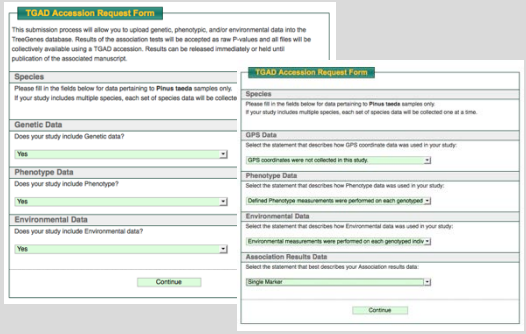
Date	Accession	Paper Title	Species	Data Statistics	Data Files
8/5/2011	TGDR001	Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem.	Populus trichocarpa	Total Sites: 1 Total Samples: 480 Total Genotypes: 419520 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 874 Total cpSSR Markers: 0 Total SSR Markers: 0 Total Phenotypes: 1344 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Covariate Data (Population Structure) Genotype Data (SNP) GPS Data Haplotype Data Phenotype Data Phenotype Definitions
9/25/2012	TGDR002	Astonishingly low genetic variation in Quercus acutissima, an important tree species in Satoyama, a traditional Japanese rural forest and agricultural landscape, revealed by chloroplast microsatellite markers	Quercus acutissima	Total Sites: 59 Total Samples: 2152 Total Genotypes: 12912 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 6 Total SSR Markers: 0 Total Phenotypes: 0 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Genotype Data (cpSSR) GPS Data Haplotype Data Supplemental Data
11/5/2012	TGDR003	Extensive selfing in an endangered population of Pinus parviflora var. parviflora (Pinaceae) in the Boso Hills, Japan	Pinus parviflora	Total Sites: 2 Total Samples: 116 Total Genotypes: 464 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 0 Total SSR Markers: 4 Total Phenotypes: 0 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Genotype Data (SSR) GPS Data Supplemental Data Supplemental Data Supplemental Data Supplemental Data
11/14/2012	TGDR004	Pollen dispersal and fine-scale spatial genetic structure of Dryobalanops lanceolata in a Bornean rain forest	Dryobalanops lanceolata	Total Sites: 1 Total Samples: 858 Total Genotypes: 13728 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 0	Environmental Metric Data Environmental Metric Definitions Genotype Data (SSR)

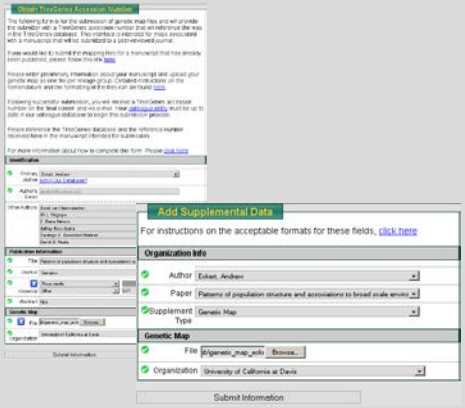
TreeGenes Database: Data Sources

treegenesdb.org

User submissions

Genetic maps, association or population studies





Obtain TGDR accession number!

TreeGenes Data Repository

A listing of data submissions is displayed below.
To submit data to TreeGenes, click here.

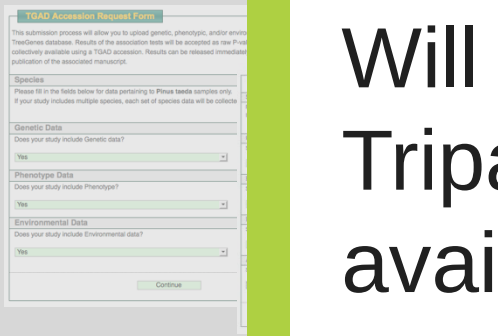
Date	Accession	Paper Title	Species	Data Statistics	Data Files
8/5/2011	TGDR001	Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem.	Populus trichocarpa	Total Sites: 1 Total Samples: 480 Total Genotypes: 419520 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 874 Total cpSSR Markers: 0 Total SSR Markers: 0 Total Phenotypes: 1344 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Covariate Data (Population Structure) Genotype Data (SNP) GPS Data Haplotype Data Phenotype Data Phenotype Definitions
9/25/2012	TGDR002	Astonishingly low genetic variation in Quercus acutissima, an important tree species in Satoyama, a traditional Japanese rural forest and agricultural landscape, revealed by chloroplast microsatellite markers	Quercus acutissima	Total Sites: 59 Total Samples: 2152 Total Genotypes: 12912 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 6 Total SSR Markers: 0 Total Phenotypes: 0 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Genotype Data (cpSSR) GPS Data Haplotype Data Supplemental Data
11/5/2012	TGDR003	Extensive selfing in an endangered population of Pinus parviflora var. parviflora (Pinaceae) in the Boso Hills, Japan	Pinus parviflora	Total Sites: 2 Total Samples: 116 Total Genotypes: 464 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 0 Total SSR Markers: 4 Total Phenotypes: 0 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Genotype Data (SSR) GPS Data Supplemental Data Supplemental Data Supplemental Data Supplemental Data
11/14/2012	TGDR004	Pollen dispersal and fine-scale spatial genetic structure of Dryobalanops lanceolata in a Bornean rain forest	Dryobalanops lanceolata	Total Sites: 1 Total Samples: 858 Total Genotypes: 13728 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 0	Environmental Metric Data Environmental Metric Definitions Genotype Data (SSR)

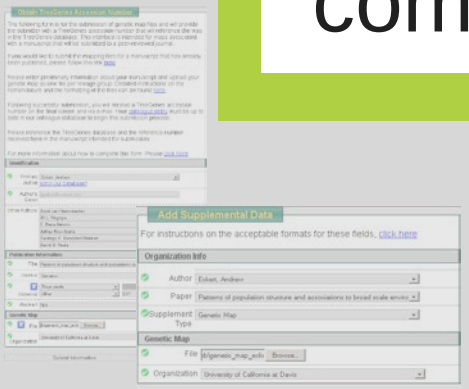
TreeGenes Database: Data Sources

treegenesdb.org

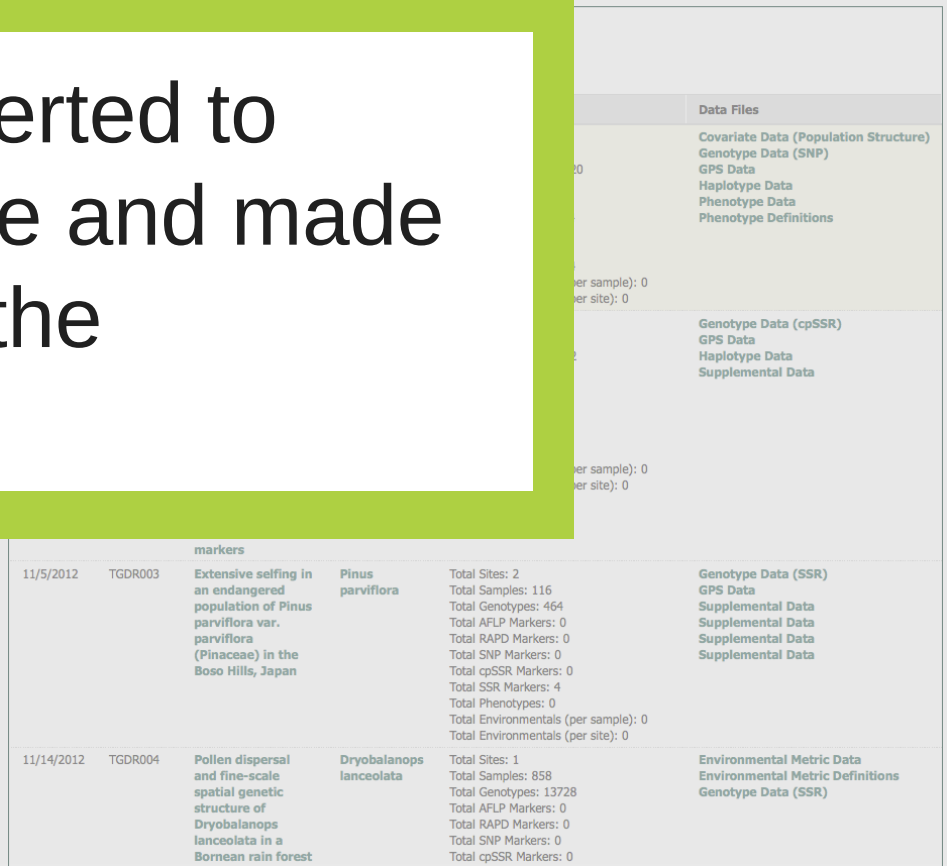
User submissions

Genetic maps, association or population studies





Will be converted to Tripal module and made available to the community



					Data Files
					Covariate Data (Population Structure) Genotype Data (SNP) GPS Data Haplotype Data Phenotype Data Phenotype Definitions
					Genotype Data (cpSSR) GPS Data Haplotype Data Supplemental Data
		markers			
11/5/2012	TGDR003	Extensive selfing in an endangered population of Pinus parviflora var. parviflora (Pinaceae) in the Boso Hills, Japan	Pinus parviflora	Total Sites: 2 Total Samples: 116 Total Genotypes: 464 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 0 Total SSR Markers: 4 Total Phenotypes: 0 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Genotype Data (SSR) GPS Data Supplemental Data Supplemental Data Supplemental Data Supplemental Data
11/14/2012	TGDR004	Pollen dispersal and fine-scale spatial genetic structure of Dryobalanops lanceolata in a Bornean rain forest	Dryobalanops lanceolata	Total Sites: 1 Total Samples: 858 Total Genotypes: 13728 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0 Total cpSSR Markers: 0	Environmental Metric Data Environmental Metric Definitions Genotype Data (SSR)

TreeGenes Database: Data Access

Tools

- Existing viewers
- Custom development

TreeGenes Database: Interfaces

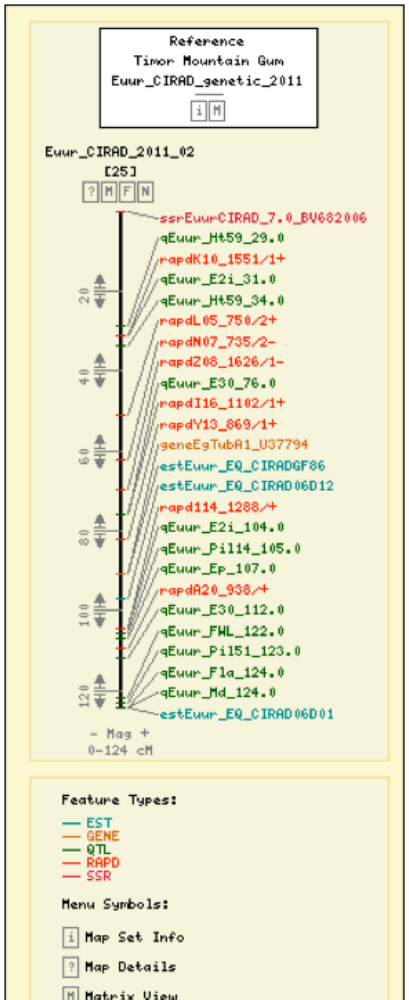
treegenesdb.org

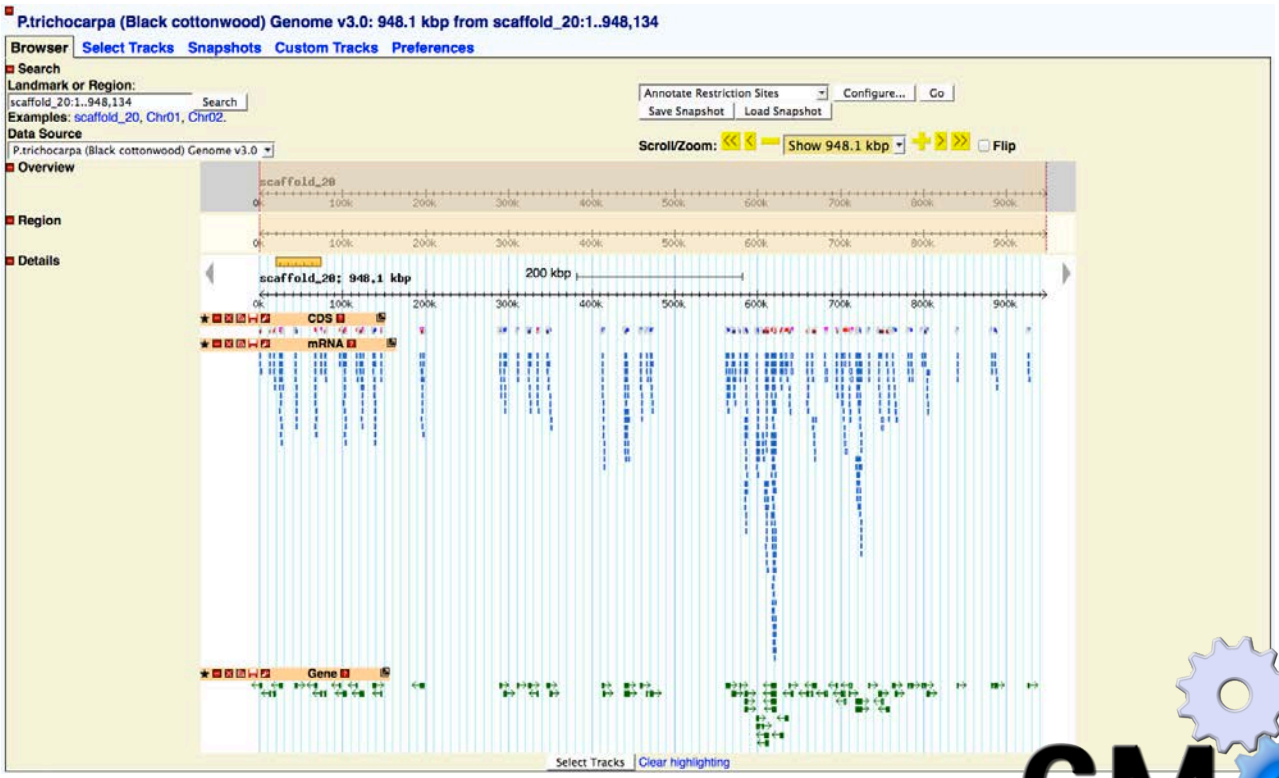
CMap

GBrowse

JBrowse

Web Apollo



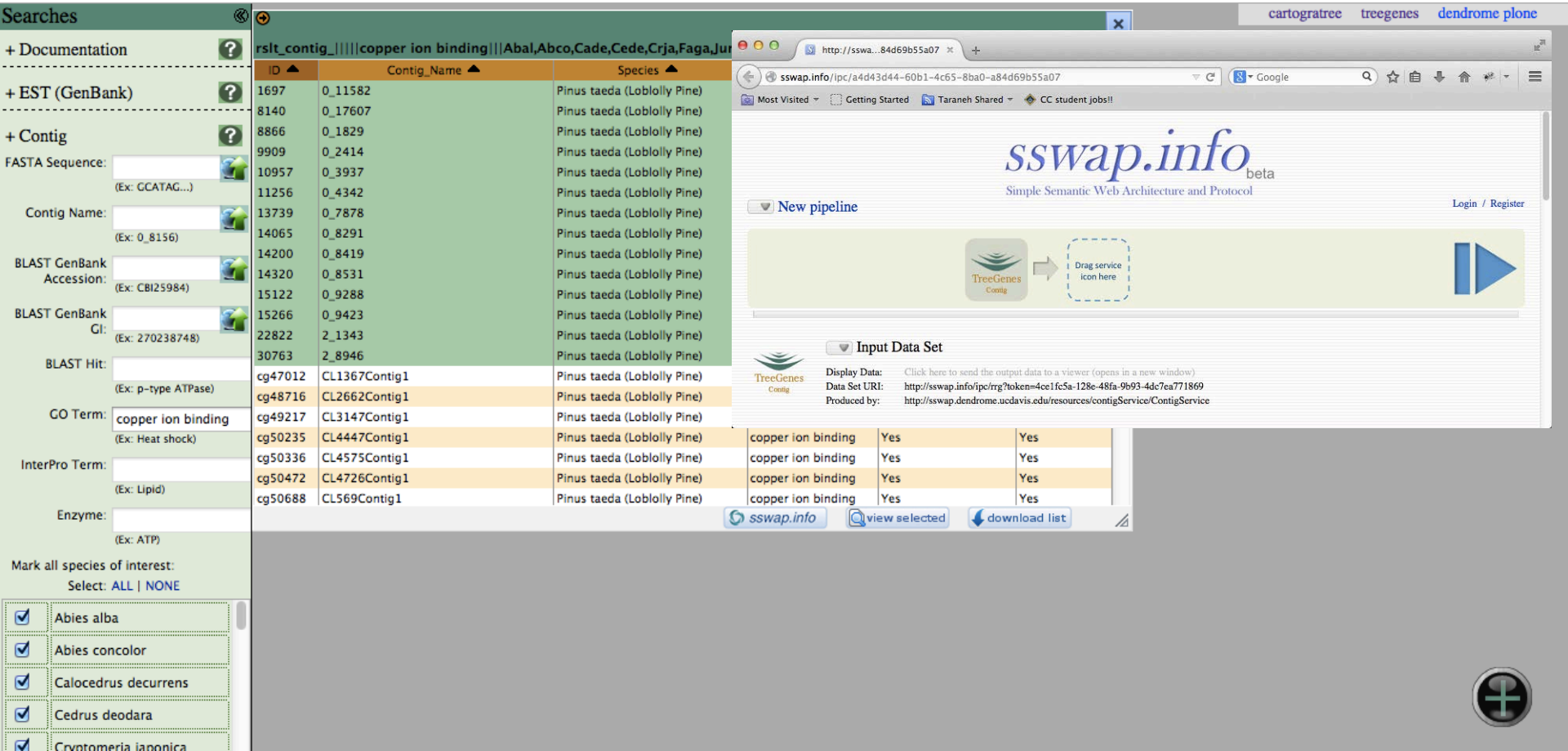


GMOD

TreeGenes Database: Interfaces

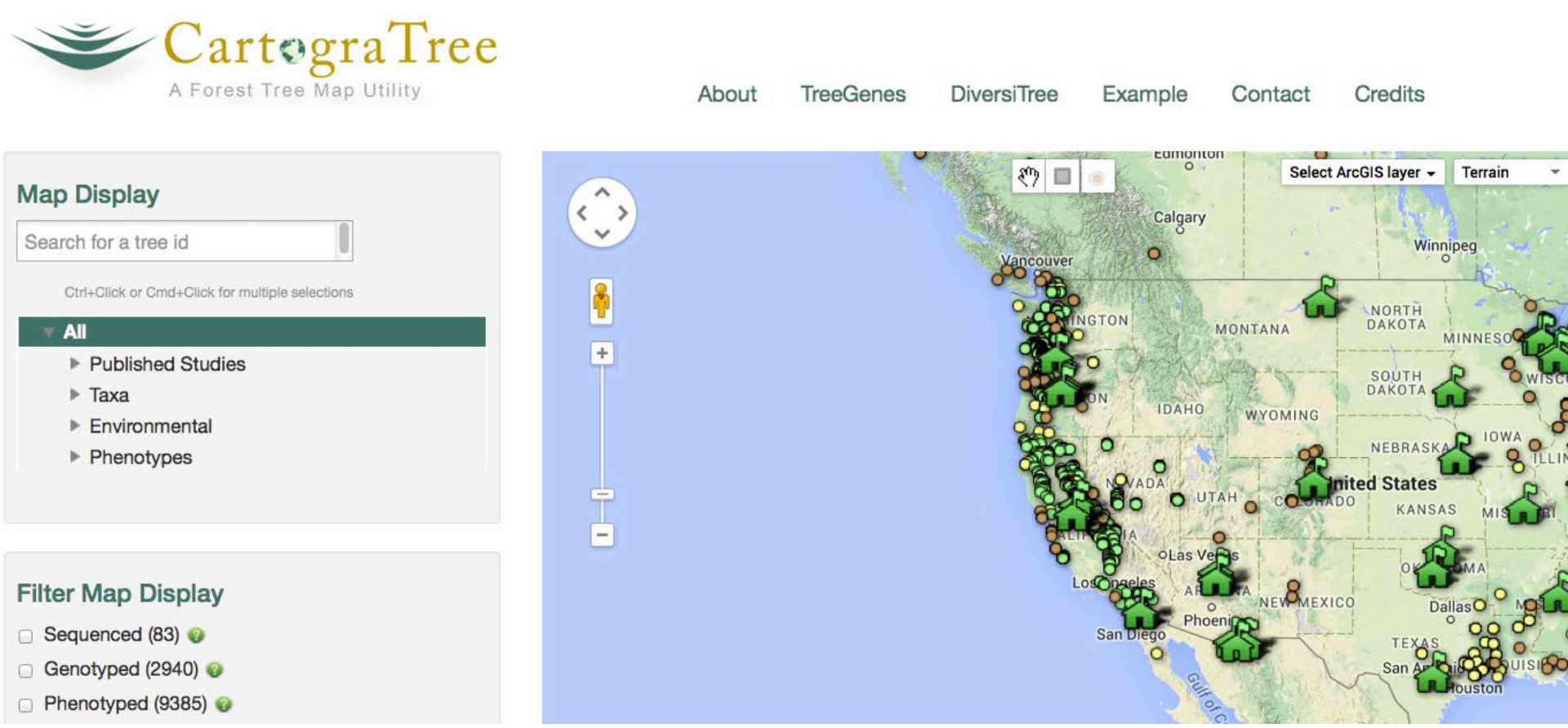
treegenesdb.org

DiversiTree – Bulk retrieval of resequencing data, genotypes, and phenotypes



TreeGenes Database: Interfaces

treegenesdb.org



- Providing context to geo-referenced data
- Data from TreeGenes, WorldClim, Ameriflux, TRY-db

TreeGenes Database:
Transition to Tripal

Transition to Tripal and Chado

Challenges

- Ontologies
- Standardizing for the first time
  - 25 years of custom schemas!

TreeGenes Database:
Transition to Tripal

Transition to Tripal and Chado

Advantages

- Ontologies
- Save time!
- Connect data with other databases (Tripal Gateway)
- Improve analytical capabilities (Tripal Gateway)
  - **Our contribution: analytical workflows**
- Review & clean up database

TreeGenes Database:
Transition to Tripal

Outcomes

- –Expanded datasets: access Hardwood Genomics, GDR, TreeGenes from one location
- –Expanded analytical, HPC resources
  - •Improve CartograTree capabilities
- –New Tripal module: user submission

TreeGenes Database: Team

treegenesdb.org **tg-help@ucdavis.edu**

Project Leads

Jill Wegrzyn

University of Connecticut

David Neale

Lead Database Administrator

Emily Grau **P0322**

Development Team

Steven A Demurjian Jr **P0383**

Hans Vasquez-Gross

Advising

Damian Gessler

Semantic Options

@TreeGenes

TreeGenes Database

UCONN UNIVERSITY OF CONNECTICUT

UCDAVIS PLANT SCIENCES

USDA

NSF