Scientific Databasing with TreeGenes:

Genotype, Phenotype, & Environment

treegenesdb.org

Jill Wegrzyn
Department of Ecology & Evolutionary Biology
Institute for Systems Genomics: Computational Biology Core
University of Connecticut, Storrs CT

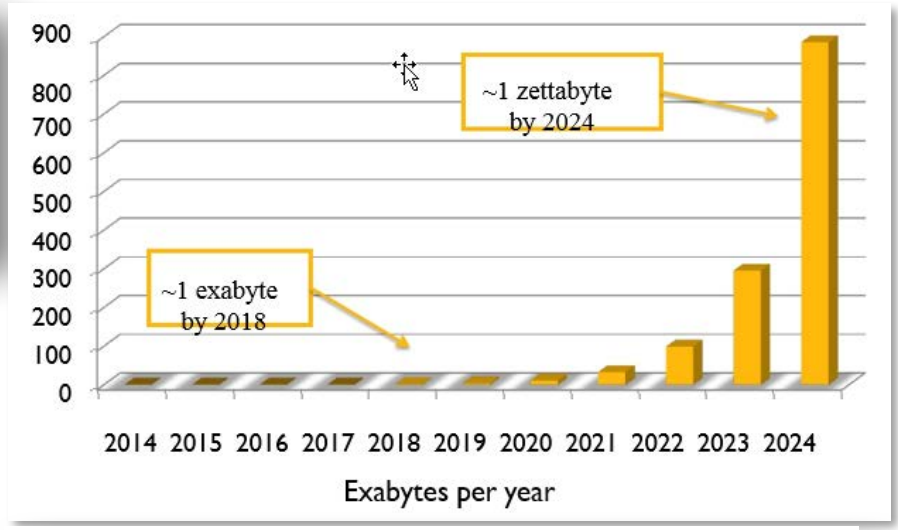
Big Data in Genomics

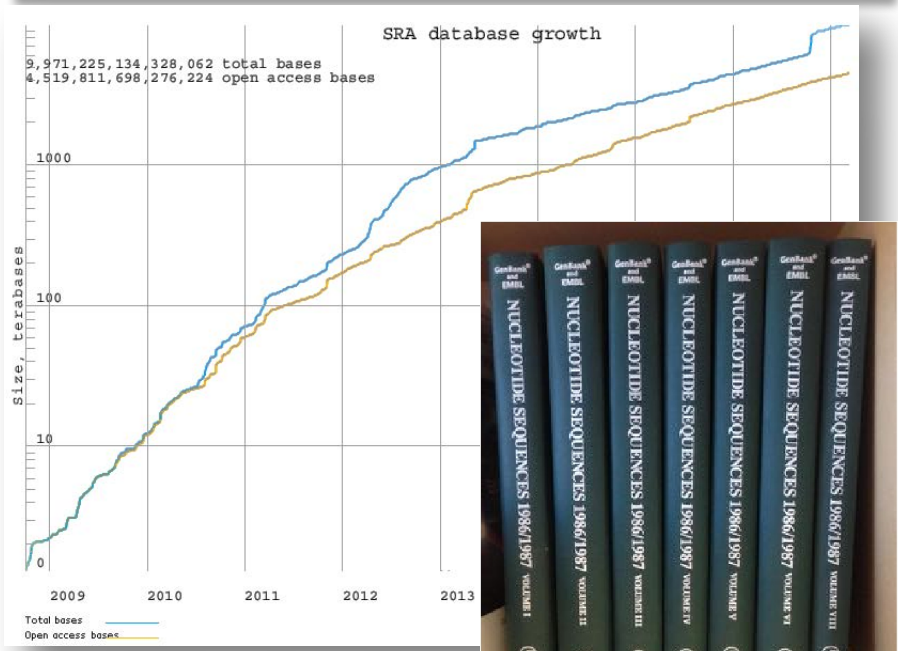
PERSPECTIVE

Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	1,000,000,000,000,000,000
Zettabyte	1,000,000,000,000,000,000,000



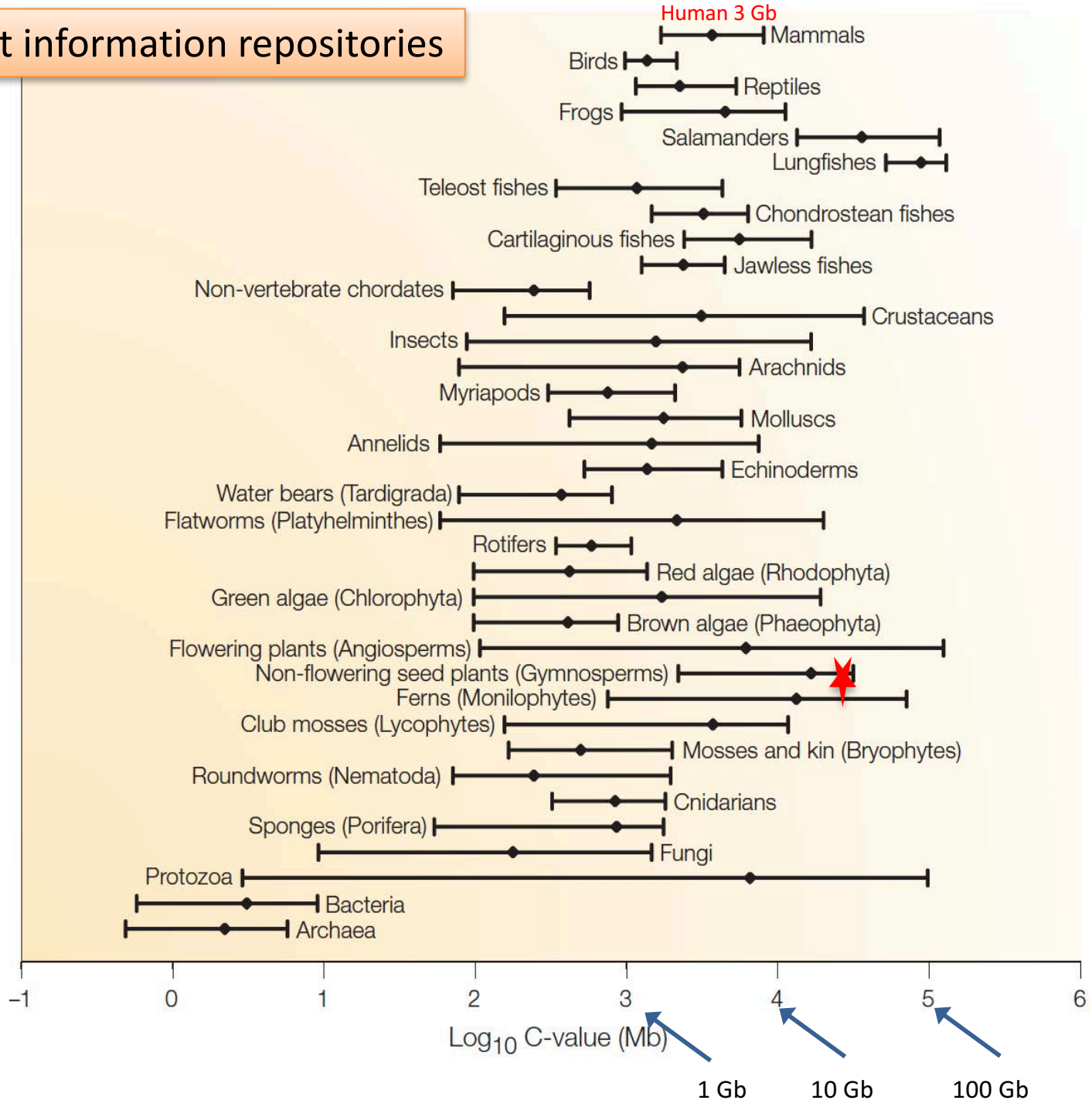


"Compared genomics with three other major generators of Big Data: ***Astronomy, YouTube, and Twitter...***Genomics is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis"

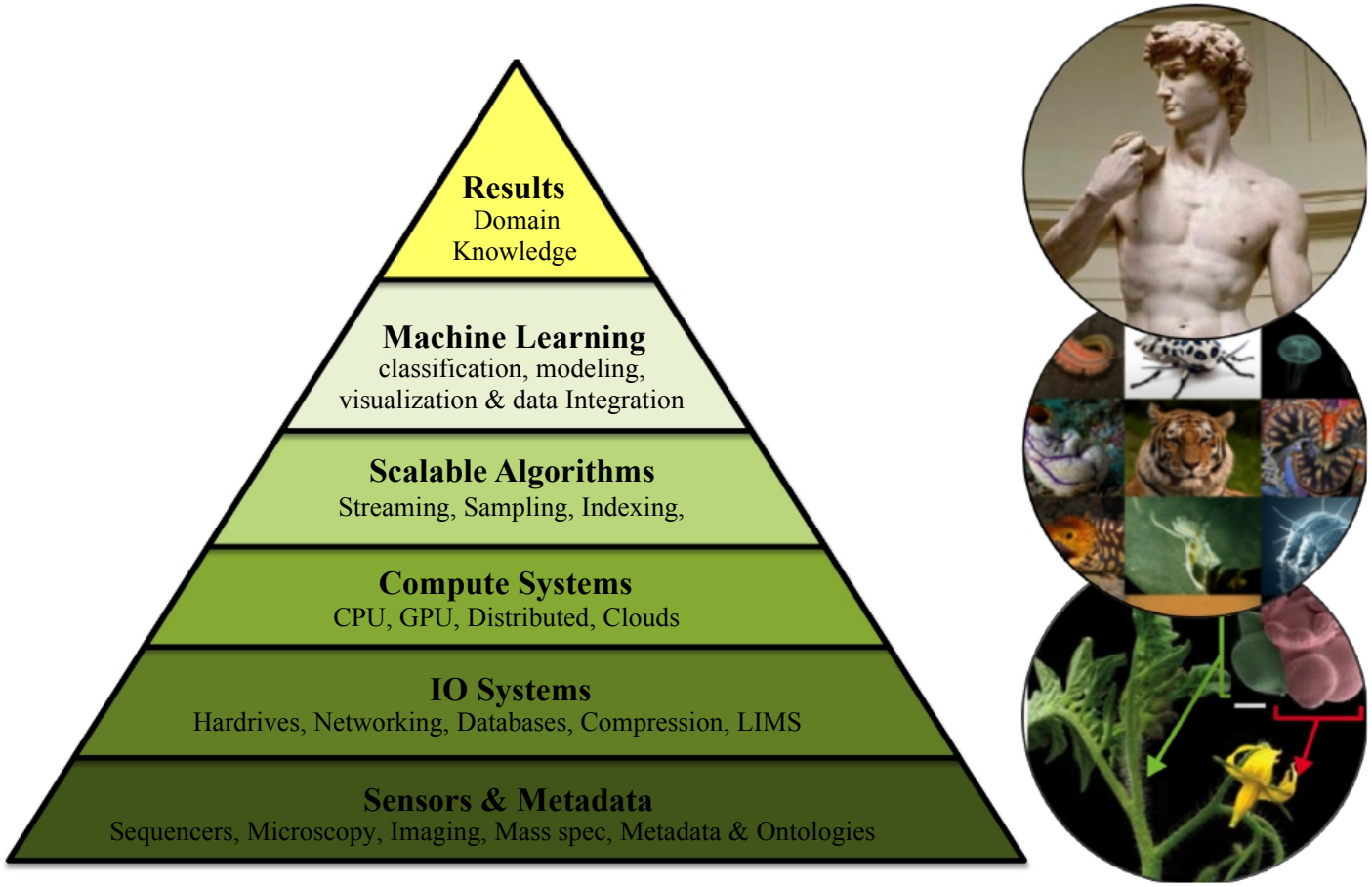
Mostly Genomic but…Proteomics, Phenomics, Metabolomics…

Genomes are vast information repositories

- Kb = 1000 bp
- Mb = $1x10^6$ bp
- Gb = $1x10^9$ bp
- Tb = $1x10^{12}$ bp
- Pb = $1x10^{15}$ bp



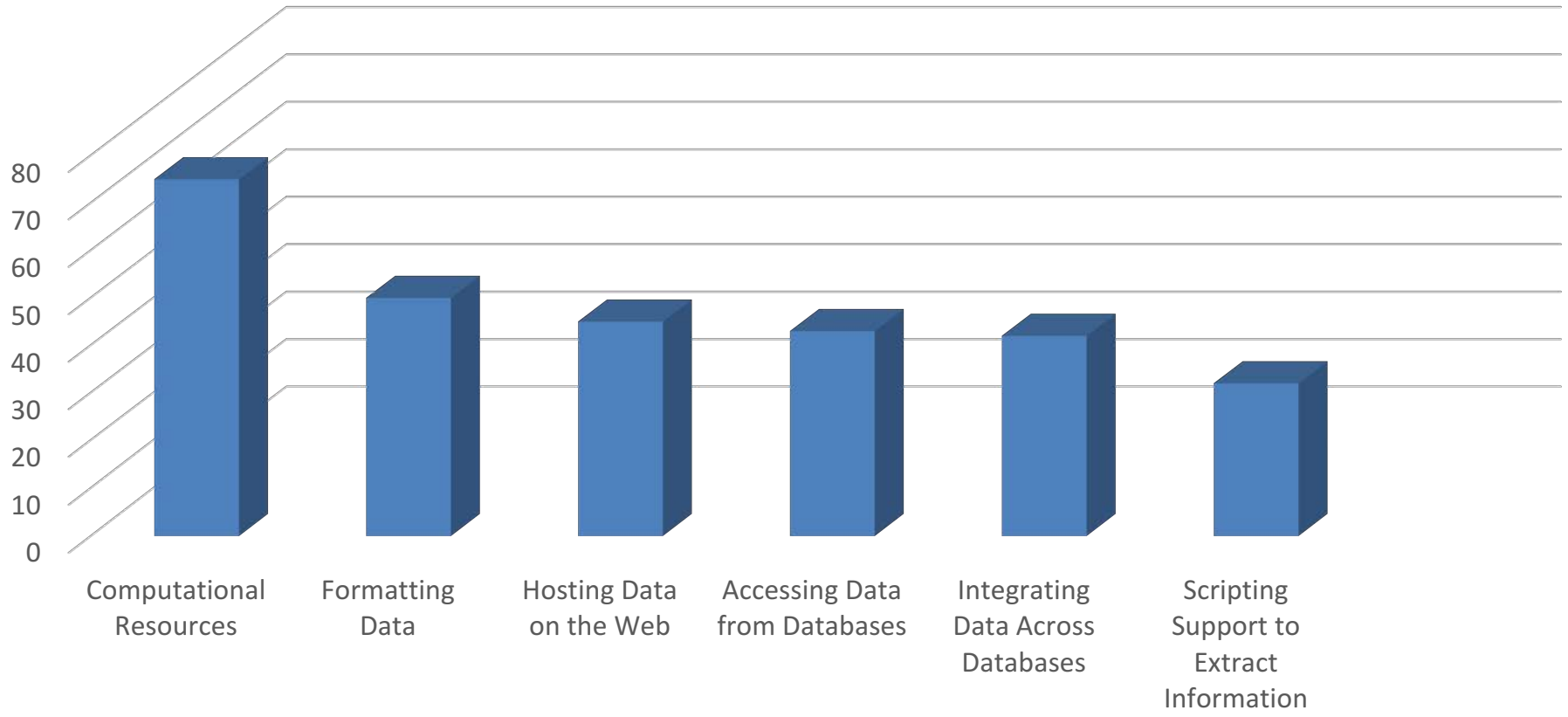
Acquiring Knowledge through Big Data



Gene Conservation of Tree Species – Banking on the Future (2016)

- Survey Conducted
  - Breeders, Geneticists, Land Managers, and Ecologists
  - 31 Questions
    - Trees (greenhouse, plots, landscape, numbers, species)
    - Data collection (devices, software)
    - Analytical tools (statistical, databases)
    - Data storage
    - Challenges
  - 283 Respondents

Gene Conservation of Tree Species – Banking on the Future (2016)



Motivation (Data Provider)

- Support next-generation data requirements for the biological database
  - Increased quantity and availability of **new data**
  - Support data **integration** across resources
  - Support **complex data analytics**
  - **Move data** efficiently

TreeGenes Database: History

treegenesdb.org

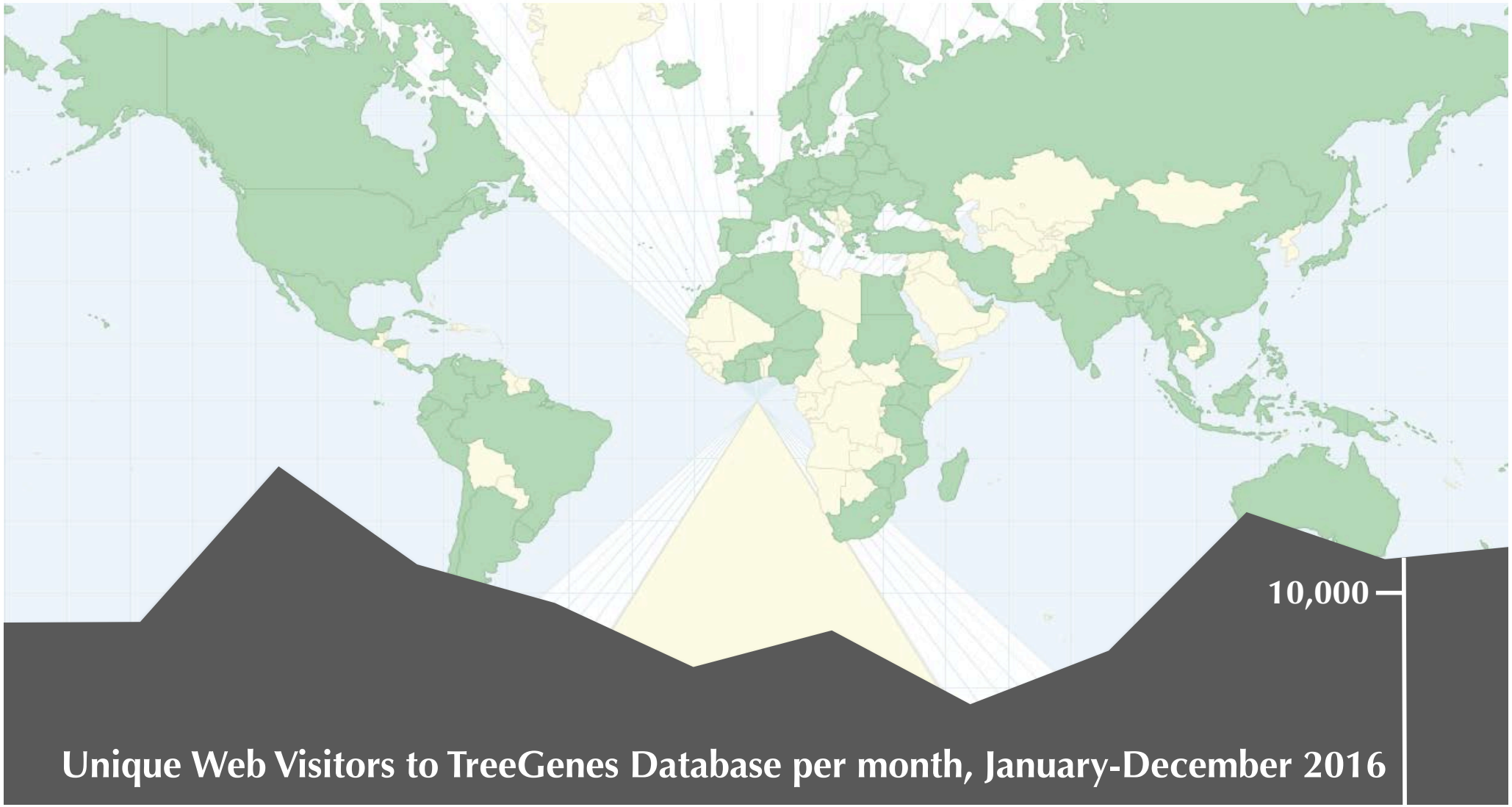
- Began to hold forest tree genetic maps and associated markers
- Expanded to other data types
  - Sequence
    - Reseqeuncing, Large-Scale Genotyping, Transcriptomics/Expression
    - Full Genome Sequences
  - Analysis and Visualization Tools
    - Ability for users to mine the data
  - Resources for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



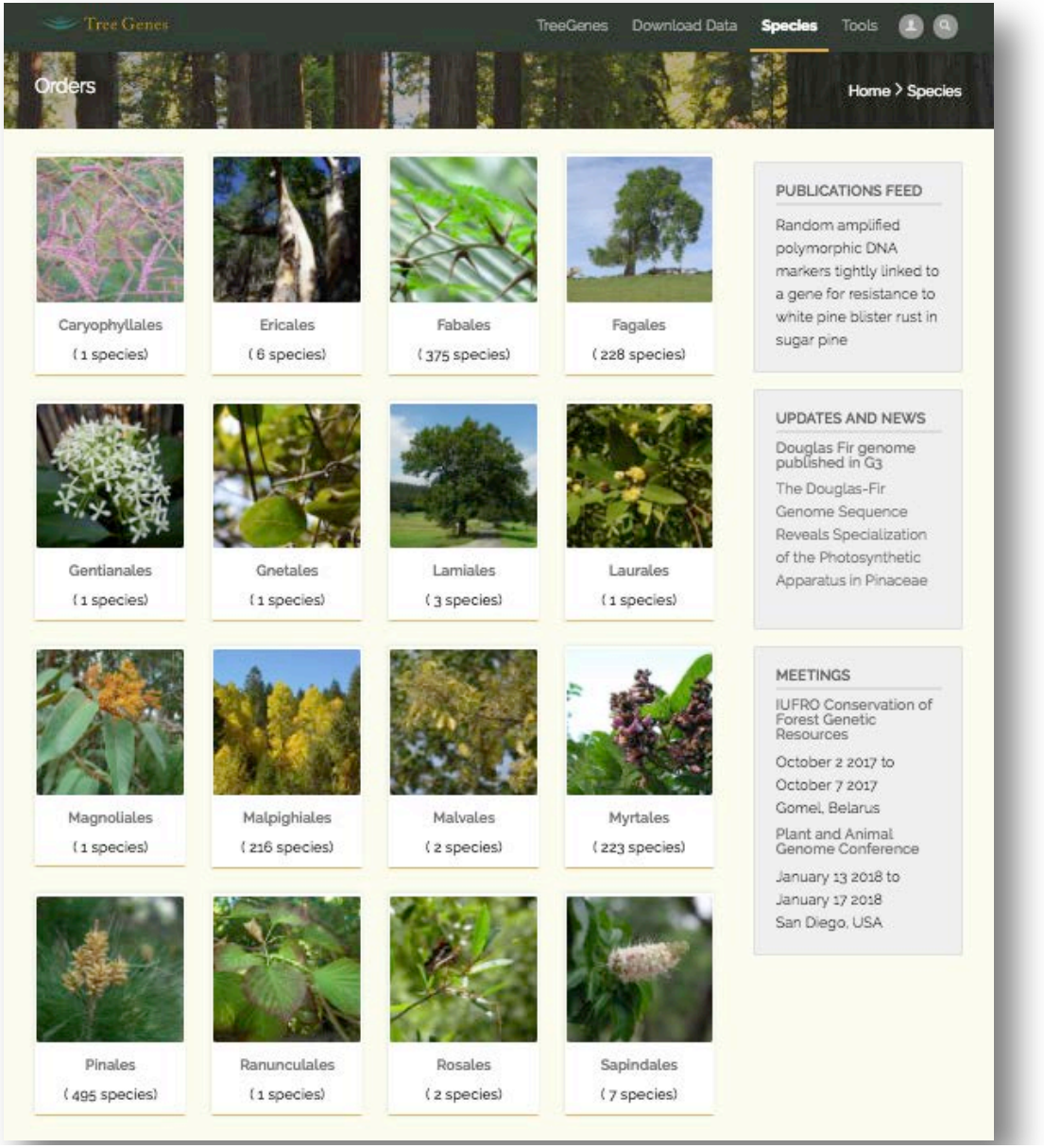
TreeGenes Database: Species

treegenesdb.org

- 1,774 species from 101 genera
  - At least one genetic artifact from each species
- Full genome sequence: 21 species
- Transcriptome/Expression resources: 4,120,817 sequences from 283 species
- 106 genetic maps from 35 species

TreeGenes Database: Species

treegenesdb.org



TreeGenes Database: Data Sources

treegenesdb.org

Primary data sources (semi-automated)

- Primary databases such as NCBI/EBI
- Appropriate data should be submitted to primary databases
- **Consistent with changing standards**
  - **Currently no repository for non-human SNPs (new!)**

User submissions

- For data and metadata not captured well by primary databases (Journals)

Project submissions

- Internal project management (private to public)

Curated Sources

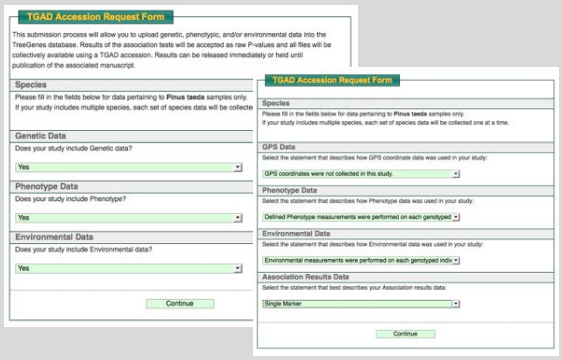
- Phytozome and PlantGDB
- PLAZA (OrthoFinder)
- TRY-DB (Phenotypes)
- Dryad (Flat files)


```
- <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.1300008
        - <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

treegenesdb.org

Data that is not collected!

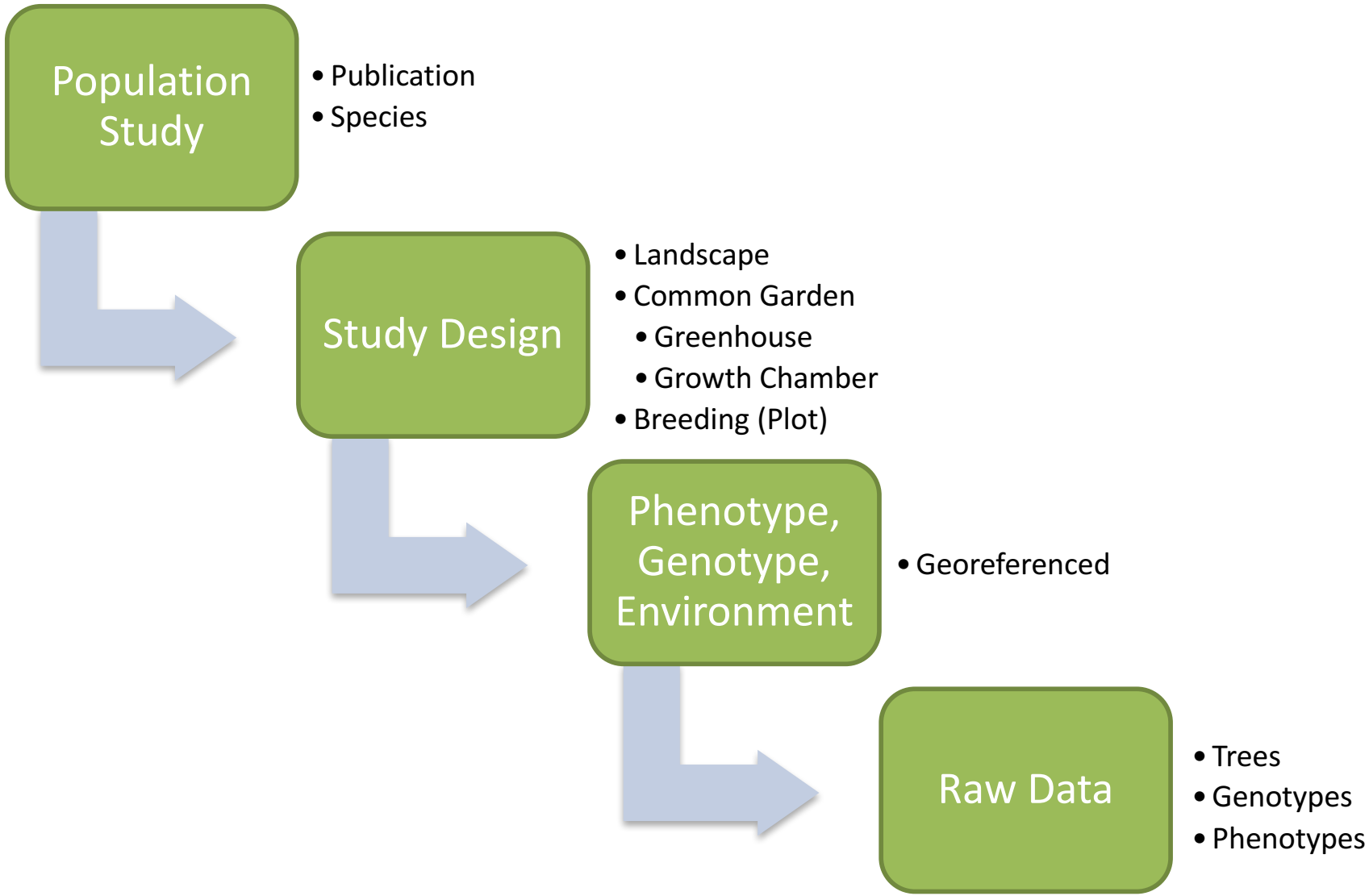


Submit **genetic maps**, **association or population study** data

Most submissions from journal requirement: Tree Genetics and Genomes, New Phytologist, and Forests

TreeGenes Database: Data Sources

treegenesdb.org

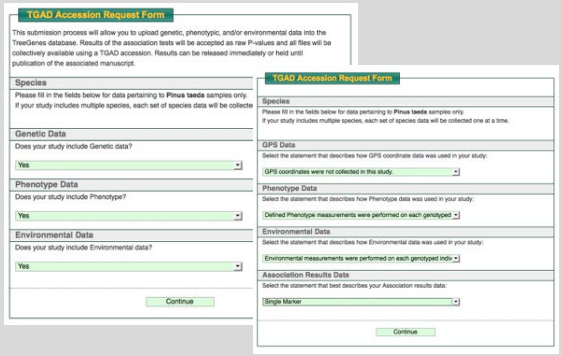


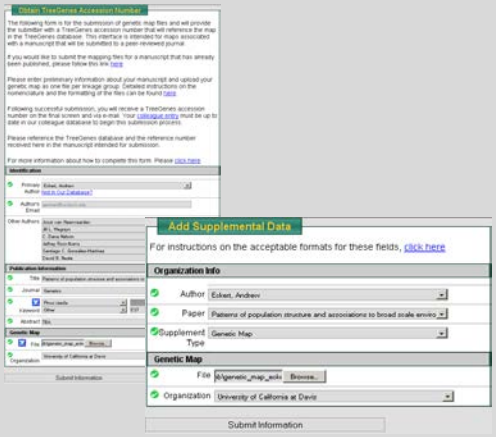
TreeGenes Database: Data Sources

treegenesdb.org

Metadata on published studies!

Genetic maps, association or population studies





TreeGenes Data Repository

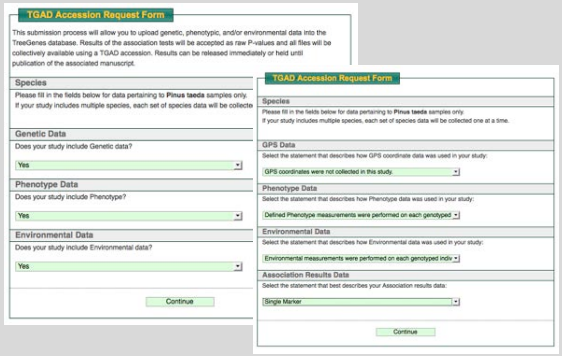
A listing of data submissions is displayed below.
To submit data to TreeGenes, click here.

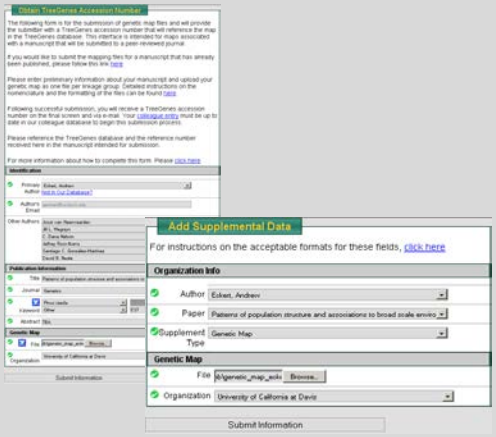
Date	Accession	Paper Title	Species	Data Statistics	Data Files
8/5/2011	TGDR001	Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem.	Populus trichocarpa	Total Sites: 1 Total Samples: 480 Total Genotypes: 419520 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

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)**

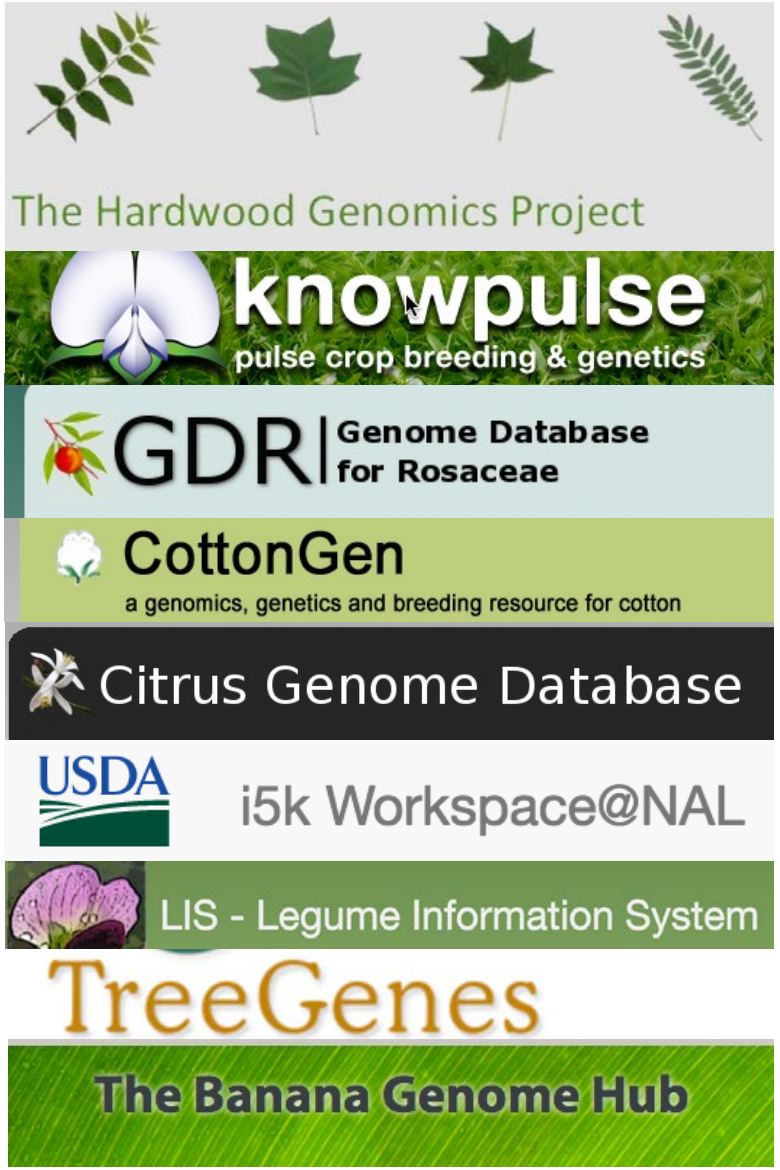
Tripal

Open source content management system (CMS) and database for biological data

Modules for genetic, genomic, and breeding data generated through a CMS and standardized schema

Benefits:

- Reduces development costs
- Provides an API for complete customization
- Uses GMOD Chado and community ontologies for standardization
- Access control for user/user groups
- Allows for sharing of extensions between sites
  - Implemented in over 30 databases!

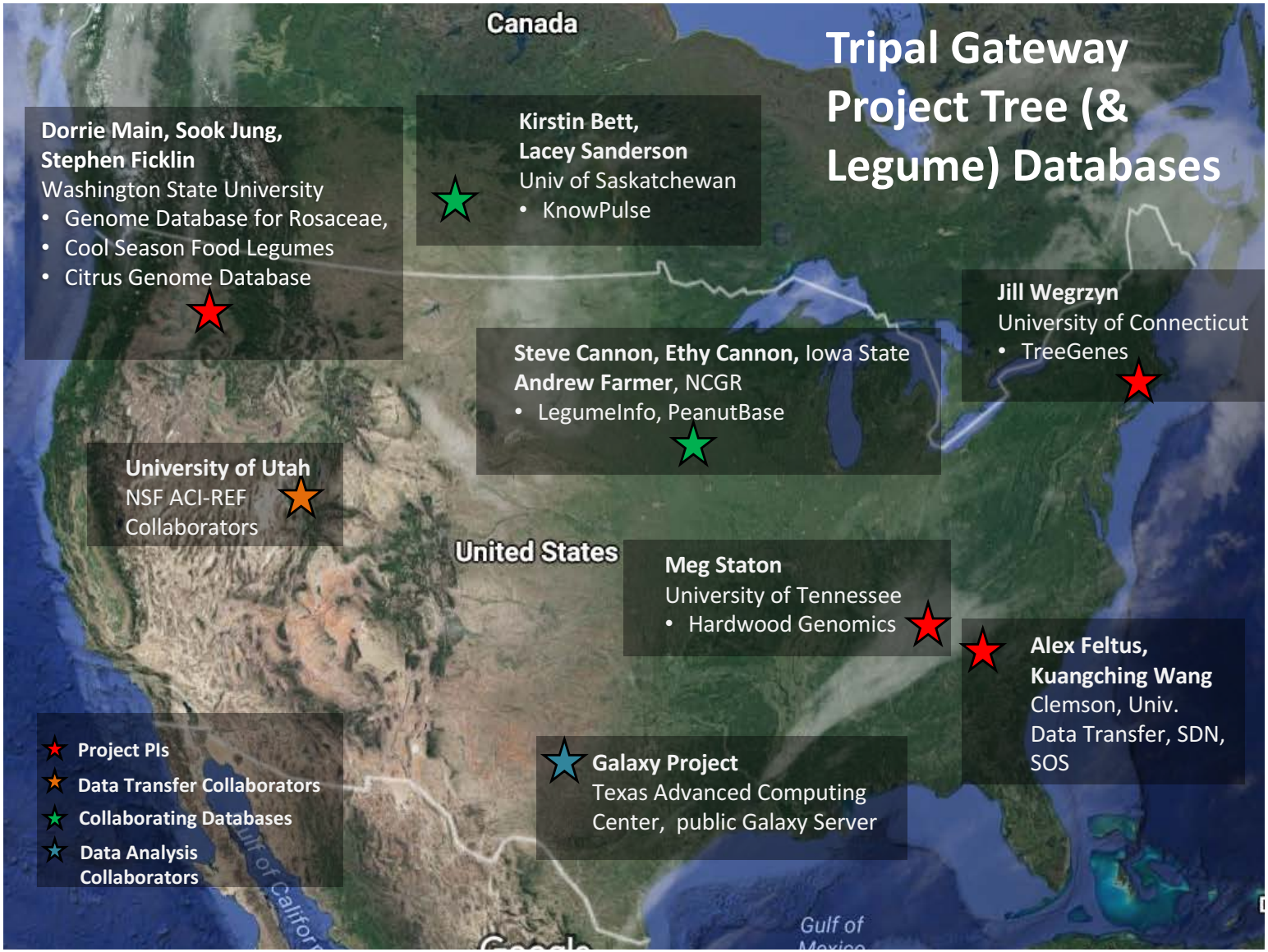


Current State of Tripal

- http://tripal.info
- Content Management System for Biological Data
- Over 100 Installations
- Current Version 2.0

Tripal Gateway Project (Data Provider)

- Support next-generation data requirements for the biological database
- **Tripal Gateway Project**
  - Increased quantity and availability of new data
  - Support **data integration** across resources (Web Services) – Tripal Exchange (v3.0)
  - Support **complex data analytics** (Integration with Galaxy API)
  - **Move data efficiently** (Software Defined Networking – Tripal Data Transfer BDSS)


Canada
Tripal Gateway Project Tree (& Legume) Databases
Dorrie Main, Sook Jung, Stephen Ficklin
Washington State University
• Genome Database for Rosaceae,
• Cool Season Food Legumes
• Citrus Genome Database
Kirstin Bett, Lacey Sanderson
Univ of Saskatchewan
• KnowPulse
Jill Wegrzyn
University of Connecticut
• TreeGenes
Steve Cannon, Ethy Cannon, Iowa State
Andrew Farmer, NCGR
• LegumeInfo, PeanutBase
University of Utah
NSF ACI-REF
Collaborators
United States
Meg Staton
University of Tennessee
• Hardwood Genomics
Alex Feltus, Kuangching Wang
Clemson, Univ.
Data Transfer, SDN, SOS
Project PIs
Data Transfer Collaborators
Collaborating Databases
Data Analysis Collaborators
Galaxy Project
Texas Advanced Computing Center, public Galaxy Server
Gulf of

TreeGenes Database: Interfaces

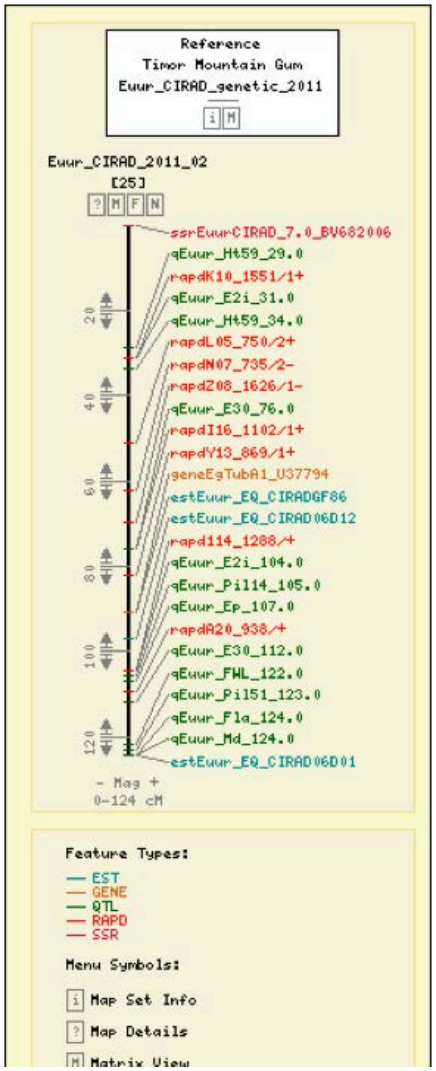
treegenesdb.org

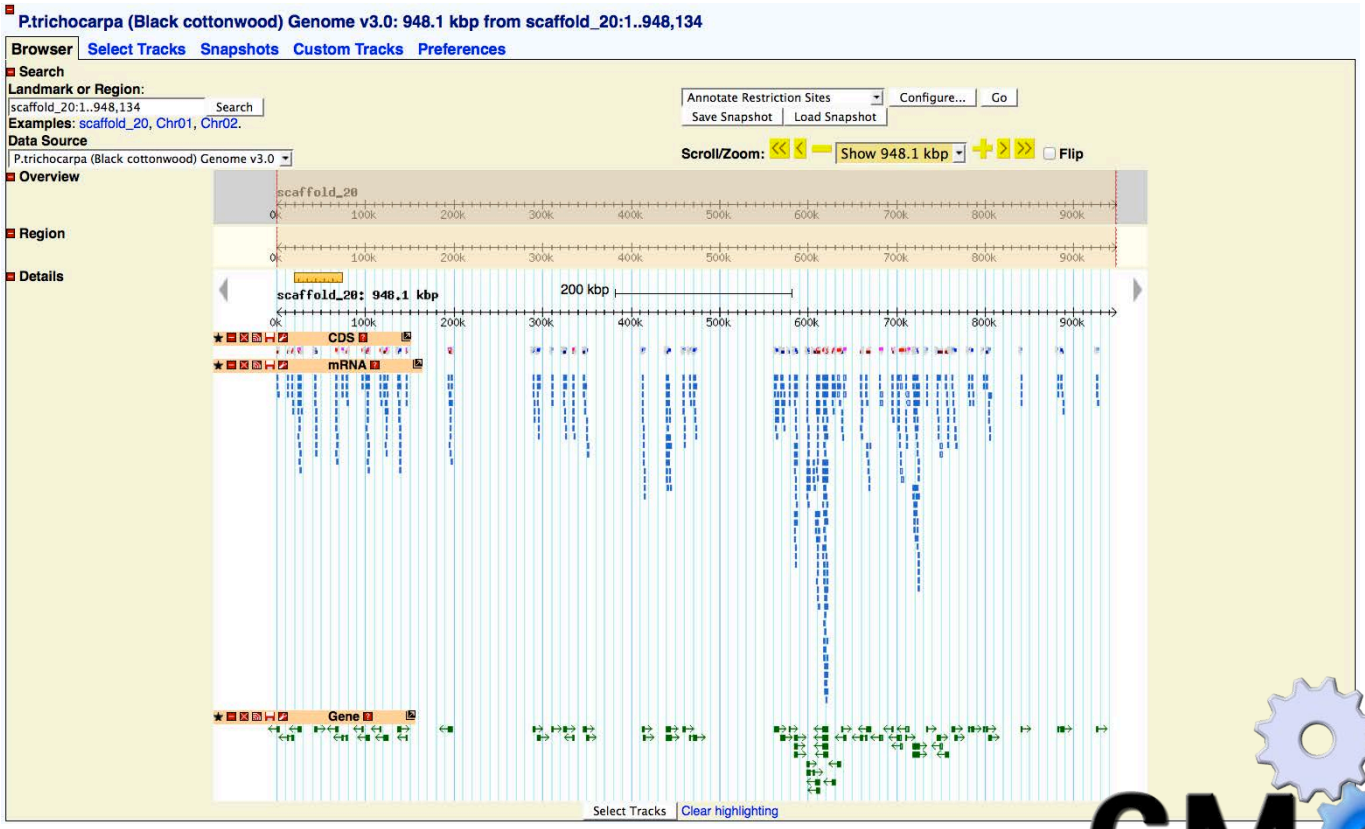
CMap

GBrowse

JBrowse

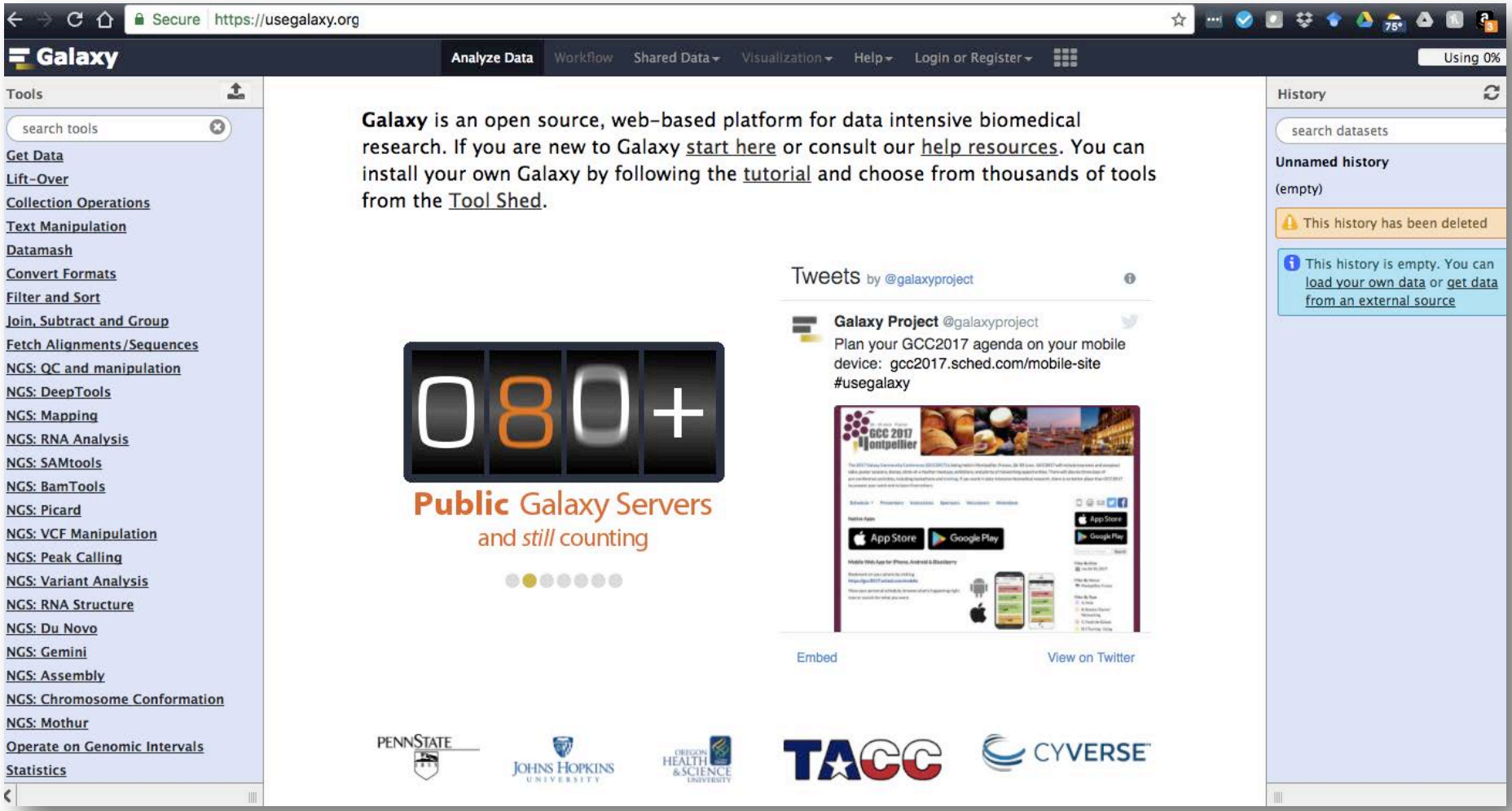
Web Apollo



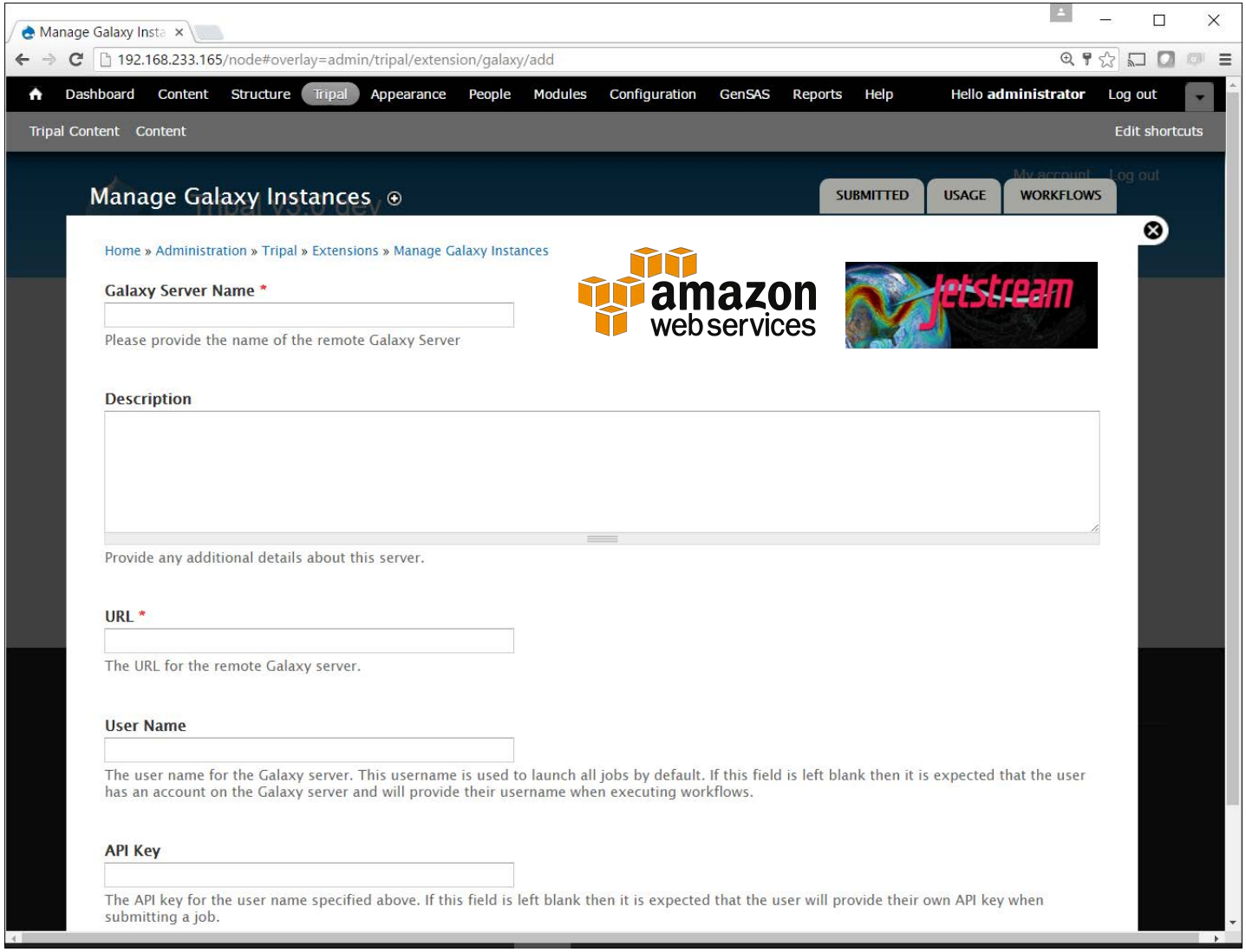


GMOD

Web-based framework (Galaxy) promotes genomics analysis



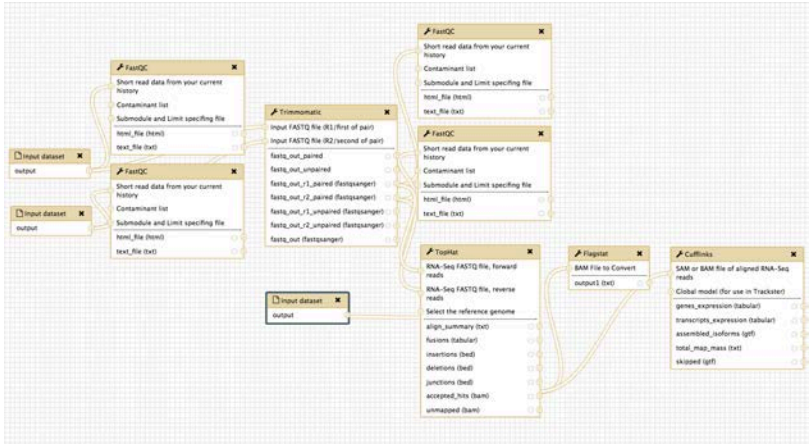
Integrating Galaxy with Tripal



Data analysis brought *to the user* via the database with Galaxy Workflows

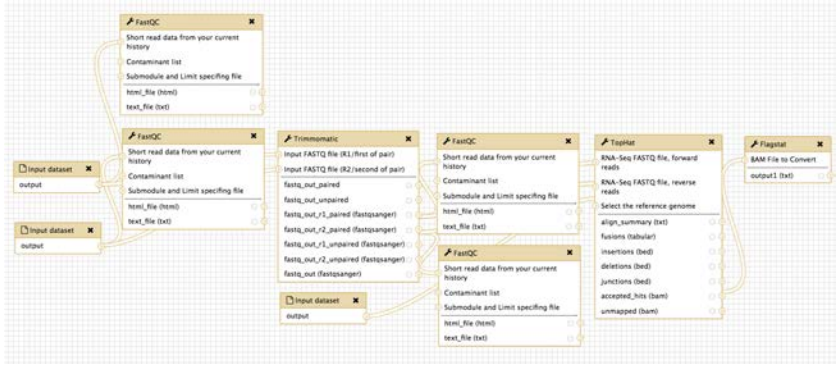
DNA Sequence Data

- Re-sequencing alignment
- Variant discovery (against the reference)
- **Variant discovery (between samples)**
- Prediction of functional genetic variants
- **Association Genetics**
- **Functional Annotation**



RNA Sequence Data

- **Transcriptome assembly**
- Alignment to a reference
- Differential Expression analysis
- Gene co-expression network construction
- MiRNA analysis



Galaxy Tool Shed

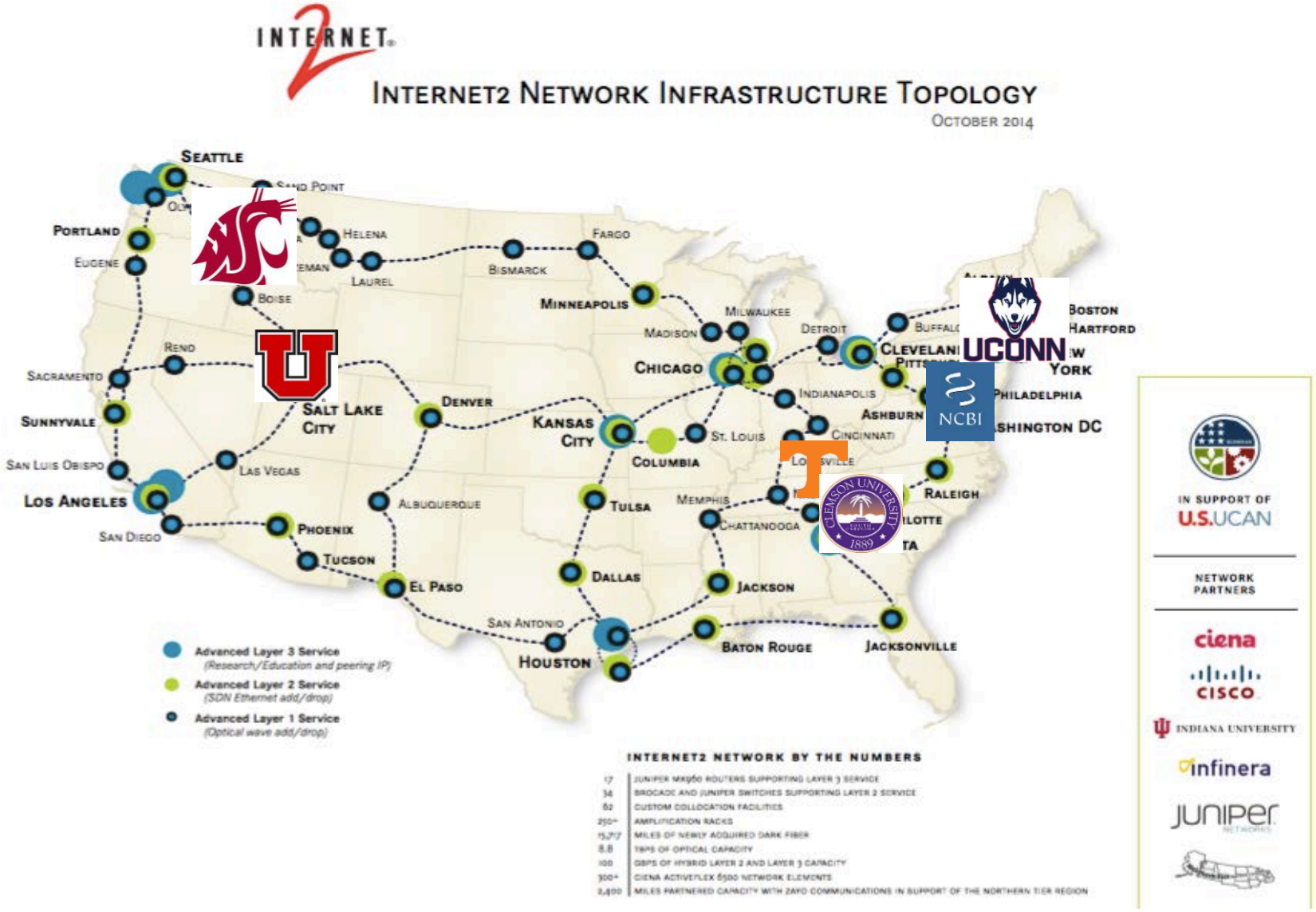
BDSS: Big Data Smart Socket

treegenesdb.org

- Smart Data Transfer
- Standalone client with a metadata repository
- First step is to build an inventory of data sources relevant to a particular user community
  - NCBI (Genbank for Raw Data)
  - Cyverse (iPlant for analytics)
  - Tripal supported websites for supporting data
- Determines optimal method for data transfer for each data source through testing
- Data transfer methodology is encoded into the metadata repository

BDSS: Moving data efficiently

treegenesdb.org



Tripal Gateway: Use Cases

Tripal Gateway:

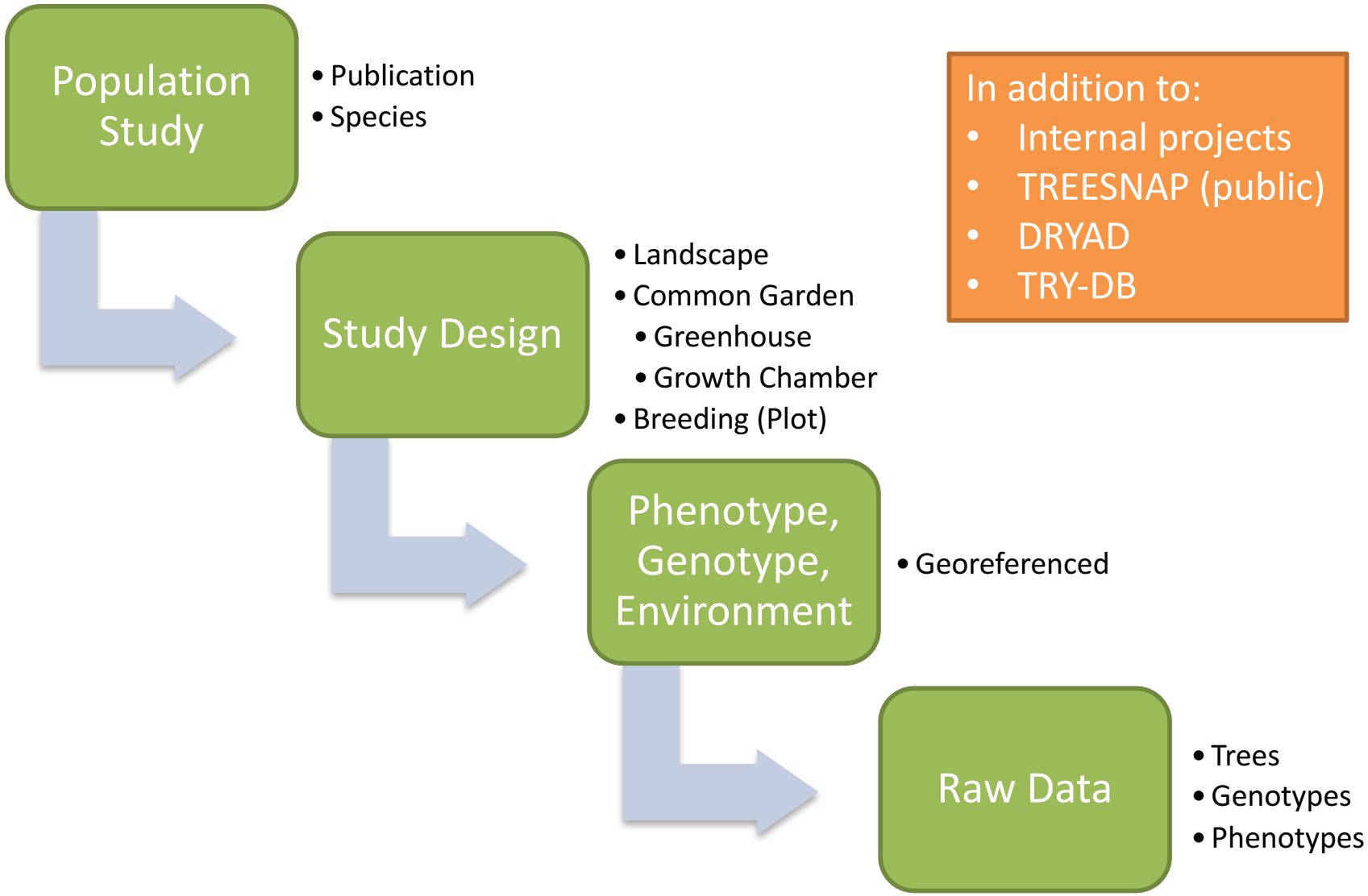
1. A user could search across community DBs for their set of SNPs interest (from a genotyping array) using Tripal Exchange.
2. The probe sequences could be gathered as a list and transferred to the user with the Data Transfer (BDSS) tool.
3. If the user prefers to use Galaxy for analysis, the transfer could load the probes into the Tripal Galaxy module and align them to a recently released genome reference
4. Basic workflow for alignment could be selected along with the appropriate target in Galaxy

Phenotype = **Genotype** + **Environment**

Provenance or Common Garden Trials	**Phenotype** X **Environmental** Associations
Marker Assisted Tree Breeding	**Genotype** X **Phenotype** Associations
Landscape Genomics	**Genotype** X **Environmental** Associations

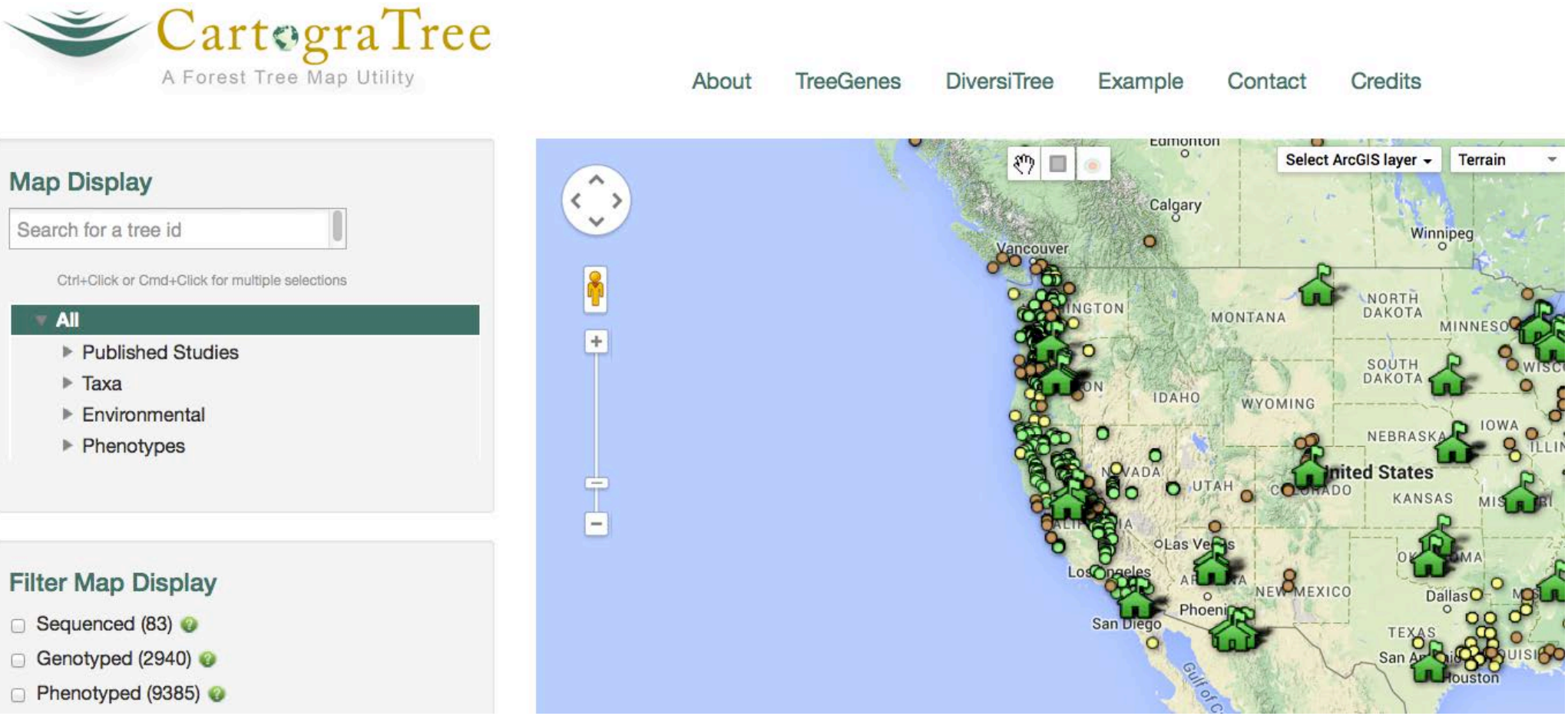
TreeGenes Database: Data Sources

treegenesdb.org



TreeGenes Database: CartograTree

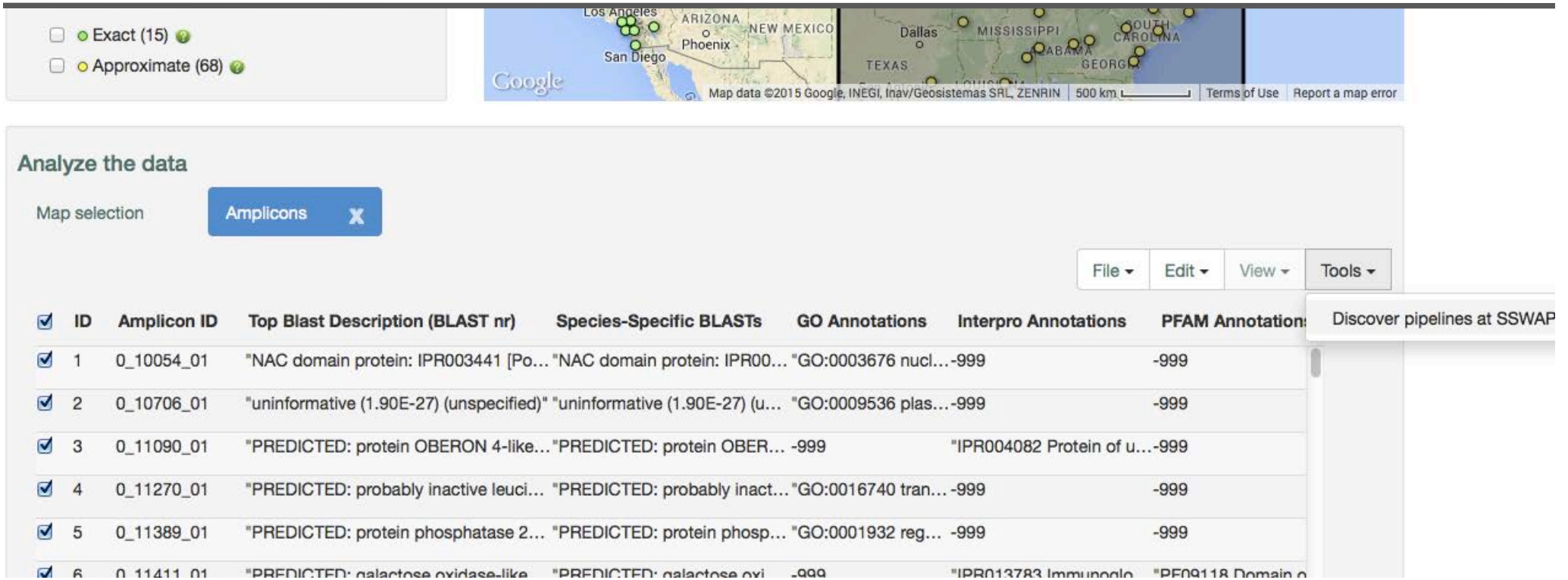
treegenesdb.org



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

TreeGenes Database: Interfaces

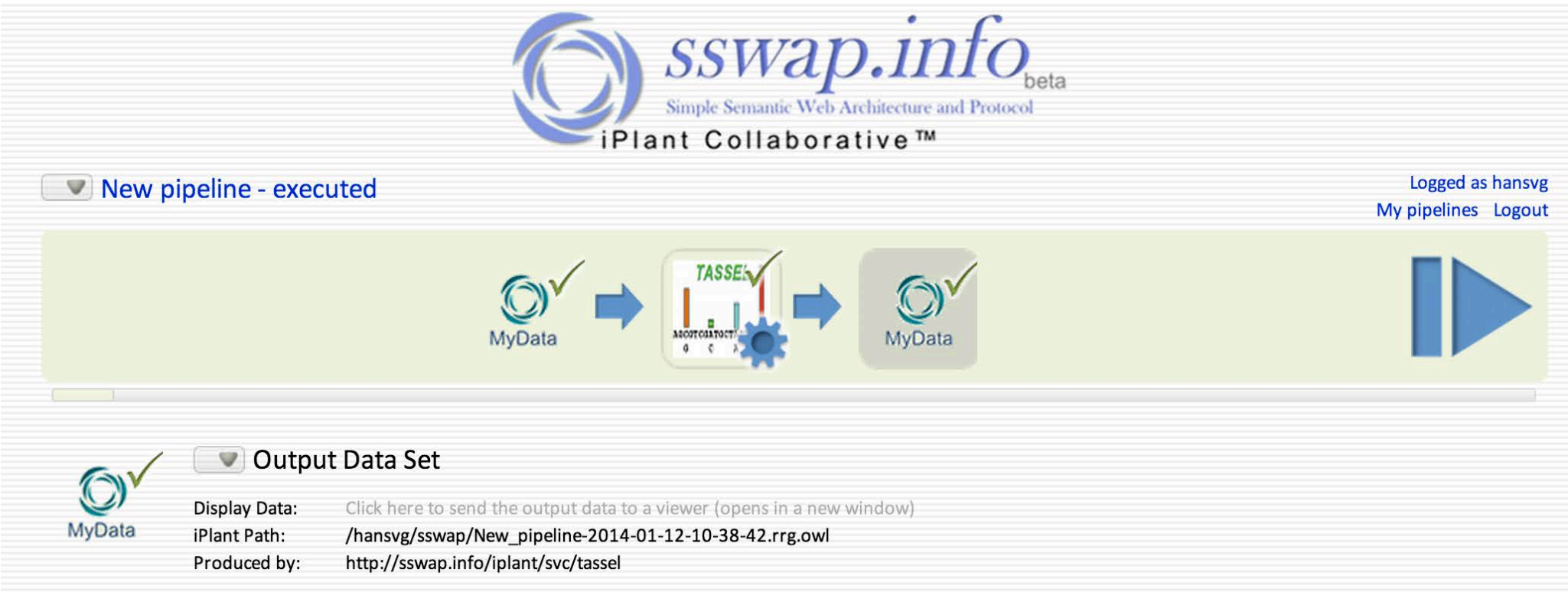
treegenesdb.org



- Retrieve genotype, phenotype, environmental, and sequence data
- Further analysis (MUSCLE, TASSEL, PAML) via SSWAP

TreeGenes Database: SSWAP

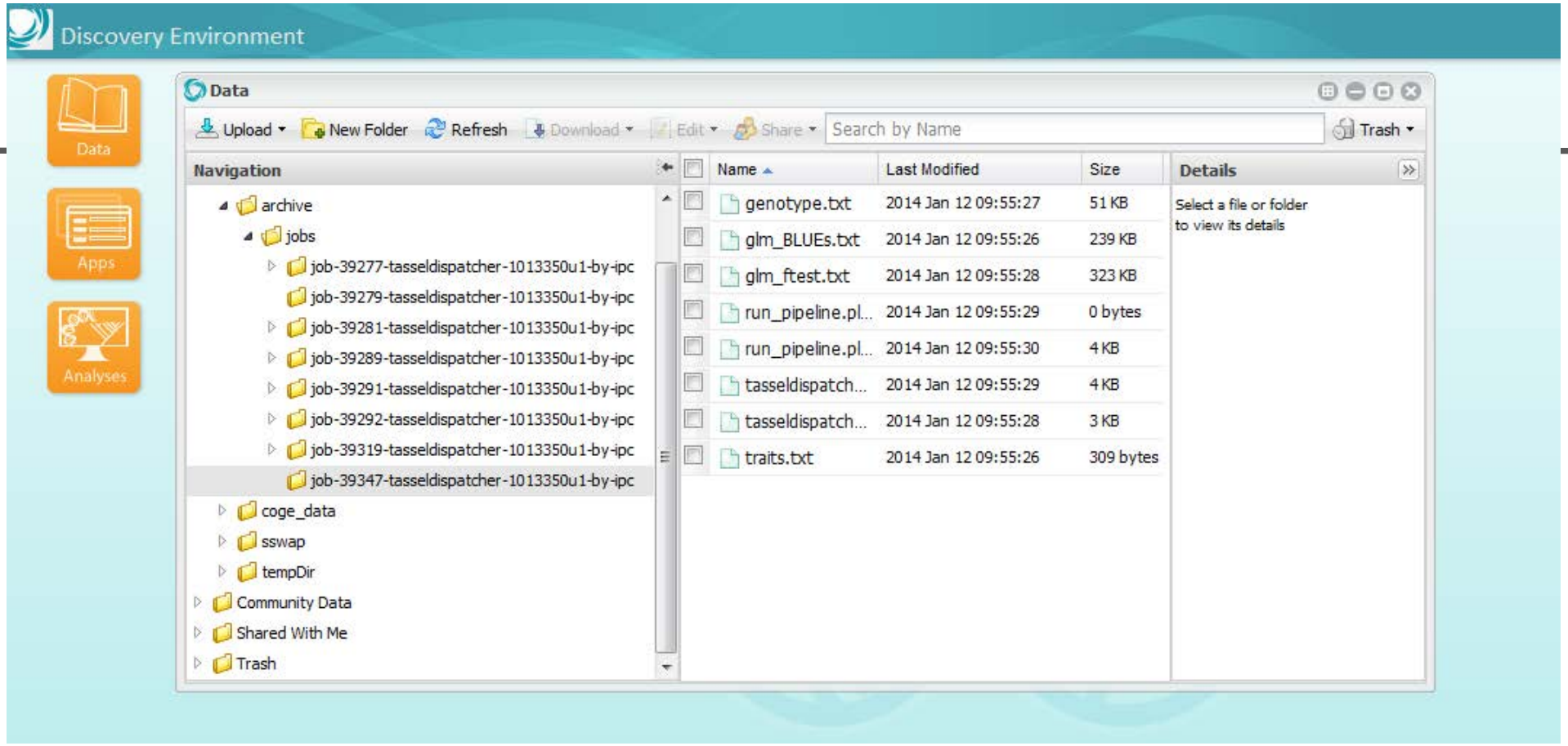
treegenesdb.org



- SSWAP "reasons" over the input data and responds with relevant applications
- Send data through pipeline with selection (parameters)

TreeGenes Database: Cyverse (TACC)

treegenesdb.org



- Connect with Cyverse Views
- Download data locally or maintain on cloud-based storage

CartograTree: Current Development

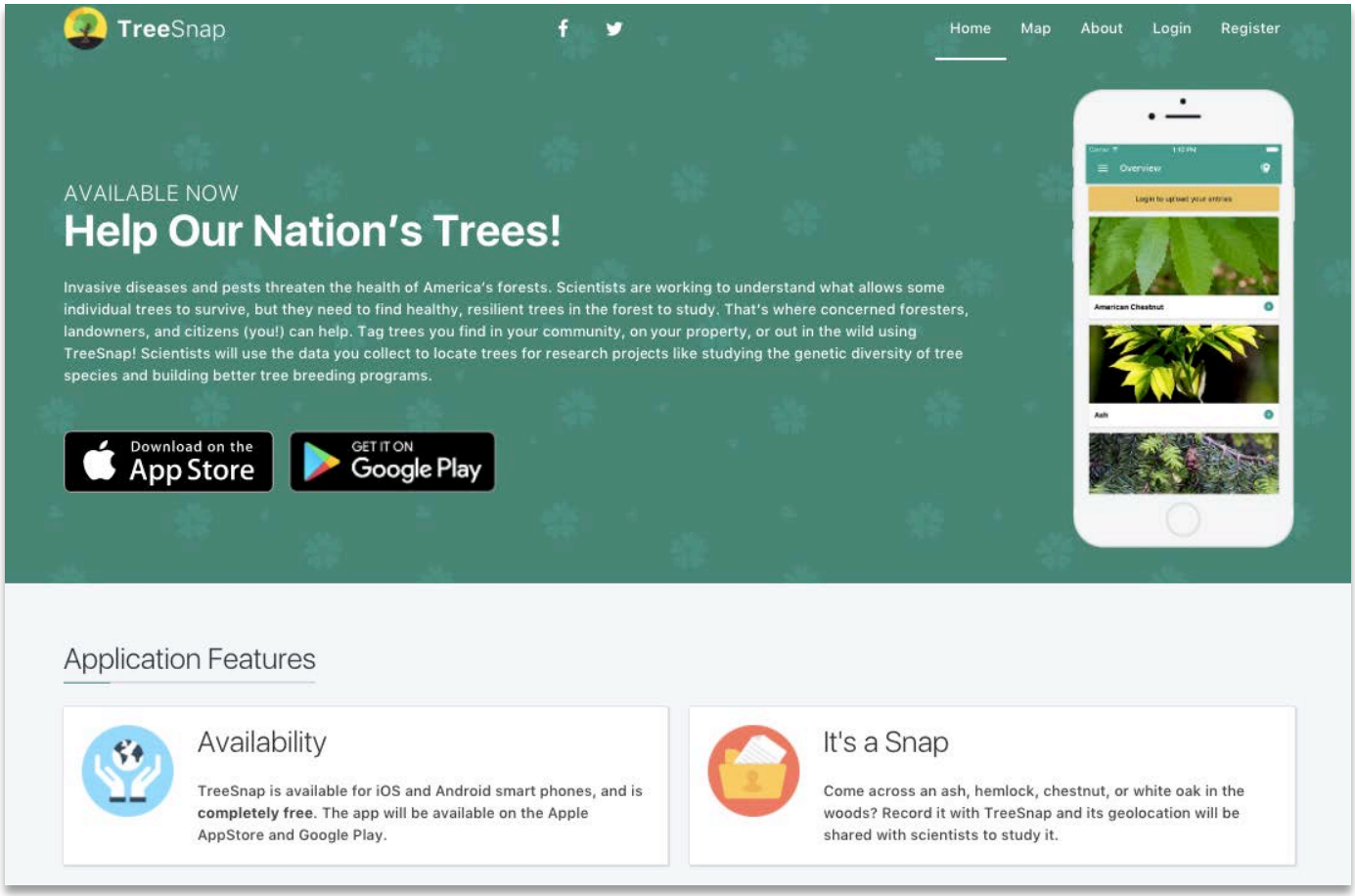
treegenesdb.org

- Flexible georeferenced tagging
  - Approximate
  - Exact
  - Obscured (radius)
- Environmental layers (Geoserver)
  - Soil
  - Fire/Drought
  - Climate models
  - LIDAR
- Integration with Tripal
  - User control of workspace
  - Ability to upload their own trees/phenotypes
- Connection with Galaxy framework
  - More analytical options (PLINK, TASSEL, MSA, PAML)
  - Intelligent workflows

CartograTree: TreeSNAP

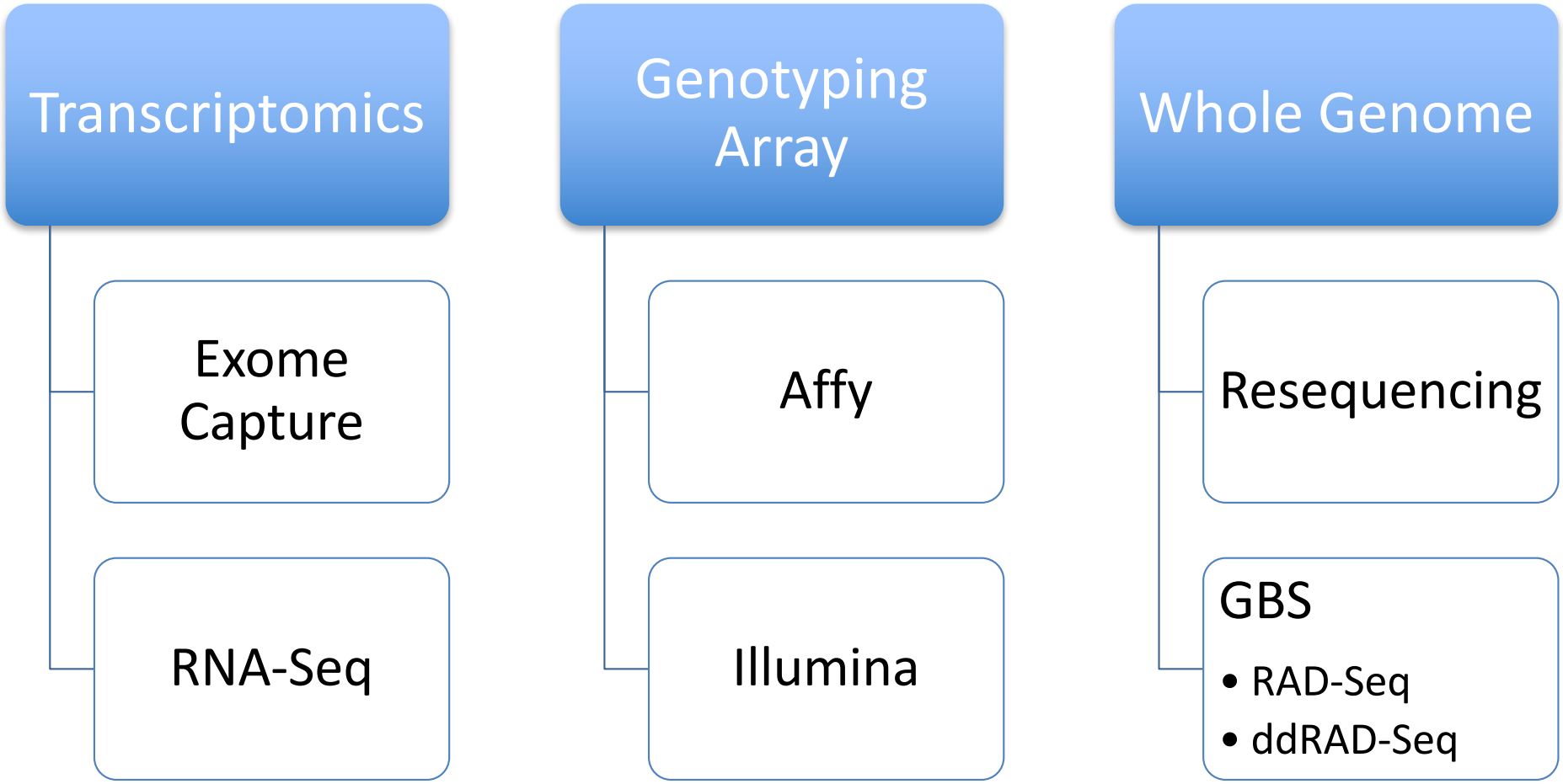
treegenesdb.org

- Validated accessions from TreeSNAP (obscured)



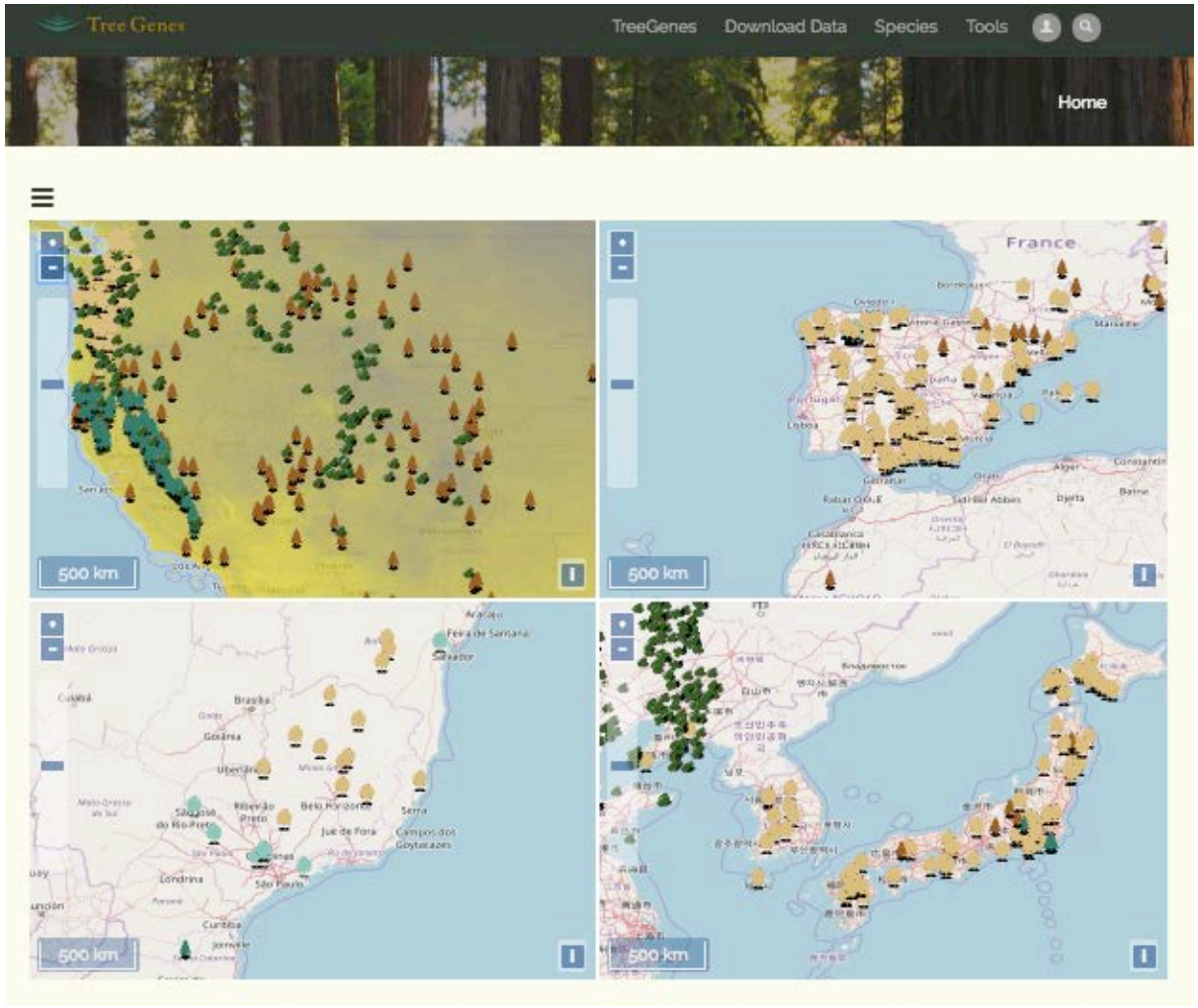
CartograTree: Galaxy Workflows

treegenesdb.org



CartograTree: Advanced Interface

treegenesdb.org



- 142 species
- 27,913 TGDR
- 17,412 Inventory
- 26,332 TRY-DB

- 815 TreeSNAP

- Release Date:
  - December 2017

TreeGenes Database: Team

treegenesdb.org **tg-help@gmail.com**

Project Leads
Jill Wegrzyn
Emily Grau
Nic Herndon

Advising
Damian Gessler

Semantic Options

Project Developers
Sean Buehler
Taylor Falk
Peter Richter
Clayton Michael

Collaborators
Stephen Ficklin (Tripal)
Alex Feltus (BDSS)
Meg Staton (HWG)
Dorrie Main (GDR)

@TreeGenes

TreeGenes Database

UCONN UNIVERSITY OF CONNECTICUT

USDA

NSF

Tripal