# 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<sup>1</sup>, Skylar Y. Lee<sup>1</sup>, Faraz Faghri<sup>2</sup>, Roy H. Campbell<sup>2</sup>, Chengxiang Zhai<sup>3</sup>, Miles J. Efron<sup>4</sup>, Ravishankar Iyer<sup>1</sup>, Michael C. Schatz<sup>5</sup>\*, Saurabh Sinha<sup>3</sup>\*, Gene E. Robinson<sup>6</sup>\*

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...





### 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

- Began to hold forest tree genetic maps and associated markers
- Expanded to other data types
  - Sequence
    - Resequencing, 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 Database: Users

treegenesdb.org

### 2,086 users from 862 organizations in 94 countries



Unique Web Visitors to TreeGenes Database per month, January-December 2016

# TreeGenes Database: Species



- 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

### 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)



treegenesdb.org

### Data that is not collected!

TGAD Accession Request Form. This submission process will allow you to upload ge	netic, phenotypic, and/or en	
TGAD Accession Request Form	netic, chencivpic, and/or en	venneantii (dati trin Pe
TGAD Accession Request Form This submission process will allow you to upload ge	metic, phenotypic, and/or en	vinnmental data into the
This submission process will allow you to upload ge	metic, phenotypic, and/or en	vinomental data into the
TreeGenes database. Results of the association tes collectively available using a TGAD accession. Resu	sts will be accepted as raw P ults can be released immedi	Avalues and all files will be ately or heid until
publication of the associated manuscript.		
Species		TGAD Accession Request Form
Please fill in the fields below for data pertaining to If your study includes multiple species, each set of	Pinus taeda samples only. species data will be collecte	Species
		Please fill in the fields below for data pertaining to Please fill in the fields amples only. If your study includes multiple species, each set of species data will be collected one at a time.
Genetic Data		
Does your study include Genetic data?		GPS Data Select the statement that describes how GPS coordinate data was used in your study:
Yes	1	GPS coordinates were not collected in this study.
Phenotype Data		Phenotype Data
Does your study include Phenotype?		Select the statement that describes how Phenotype data was used in your study:
Yes	-	Defined Phenotype measurements were performed on each genotyped -
Environmental Data		Environmental Llata Select the statement that describes how Environmental data was used in your study:
voes your study include Environmental data?		Environmental measurements were performed on each genotyped indiv -
Yes	-	Association Results Data
		Select the statement that best describes your Association results data:
	Continue	Single Marker
Resident Frankform & Access of Name 2 The Franking Waters to the Submersion of prior the Labolation with a Transform & access of the the Control of Submers and the Submersion of the the Submersion and the Submersion of the Submersion of the Submersion of Submersion of	to map time and will provide or that will reference the map ended for maps ensurether	Contra
Control for the second of	So the first and with provide the definition of the second second second encounter of the second second second second encounter of the second second second second the second second second second second the second second second second second second second second second second the second se	Griffia
Control to exceed the second on the second of the sec	the stand and provide more than a stand and provide more than a stand and a stand and a more stand and a stand a stand and a more stand and a stand a stand and a more stand and a stand a stand a more stand and a stand a stand a more stand and a stand a more stand and a stand a more stand and a	Contra
Control of the second of	In the sector of a	International Data International Soft these fields, pick more
Control of a statistical statiste statistical statistical statistical statistical statistical sta	In the first and with profiles of the second	Emerginal Data enceptable formats for these fields, <u>gick herce</u>
Control to a control to control to a co	Control of provide the second se	Iemenfal Data on the acceptable formats for these fields, give here
Control of the second of	Construction of the second sec	International Data International Store these fields, gate areas
Control of a standard a stan	A more than a set of a more than a more th	In the acceptable formats for these fields, <u>pick here</u>
Control on the section of the s	Constructions     Constructions     Construction	Emerical Data on the acceptable formats for these fields, <u>give here</u> are. Ander: are of population means and annealance to bread and and and and are of population means and annealance to bread and and and and are of population means and annealance to bread and and and and are of population means and annealance to bread and and and and are of population means and annealance to bread and and and and are of population means and annealance to bread and and and and are of population means and annealance to bread and and and and are of population means and annealance to bread and and and and are of population means and annealance to bread and and and are of population means and annealance to bread and and and are of population means and annealance to bread and and and are of population means and annealance to bread and and and are of population and annealance to bread and annealance to bread and anneal are of population and annealance to bread annealance tobread annealance to bread annealance to bread annealance tob
Balance of the second sec	Construction     Construction	
Control or contro or control or control or control or control or control or control	Annual of the set	
Control of an annual section of a secti	Construction     C	

Submit genetic maps, association or population study data

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



treegenesdb.org

### Metadata on published studies!

### Genetic maps, association or population studies

submission process will allow you to upload gene 3enes database. Results of the association tests thvely available using a TGAD accession. Result cation of the associated manuscript.	tic, phenotypic, and/or en will be accepted as raw P s can be released immed	vironmental data into the -values and all files will be allely or heid until
scles		TGAD Accession Request Form
se fill in the fields below for data pertaining to Pir	nus taeda samples only.	
ur study includes multiple species, each set of sp	ecies data will be collecte	Species Please fil in the fields below for data pertaining to Pinus taeds samples only. If your study includes multiple species, each set of species data will be collected one at a time.
netic Data		
a your study include Genetic data?		GPS Data Salest the statement that describes how OPS coordinate data was used in your study.
		GPS coordinates uses not collected in this study.
notype Data		Disanshima Data
s your study include Phenotype?		Select the statement that describes how Phenotype data was used in your study:
	-	Defined Phenotype measurements were performed on each genotyped -
vironmental Data		Environmental Data
s your study include Environmental data?		Select the statement that describes how Environmental data was used in your study:
		Environmental measurements were performed on each genotyped indiv -
	2	Association Results Data
		Select the statement that best describes your Association results data:
Co	ntinue	Single Marker •
		Continue
Figure and the transmission of the energy of the transmission of the energy of the transmission of the energy	anunopti and upped your materity and upped your restructions on page restructions on page restructions on page restorements and any office and any office and any office and any office and any office and any office any office any office any office and any office any office any office any office any any office any office any office any office any any office any office any office any office any office any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any any office any office any office any office any office any office any any office any office any office any office any office any office any office any any office any	
Other Authors Anna on American Inc.	Add Supp	lemental Data
C. Tana Maton Malay Taon Kara Tantapo C. Secular Haring Tantapo C. Secular Haring	For instructions	on the acceptable formats for these fields, <u>cack here</u>
Path dist Manualan	Organization Info	
5 The Parent of population manage and accordance in		
O D Processo al INN	Author Ea	nat, Andrew
Lawrence Caller	Paper Pa	terns of population structure and apposiations to broad scale enviro
Contract No.		anti Mari
	Supplement In	
P D to Barrow, Par. Marcon.	Supplement Ge Type	1000 May 21
The Deprese on the Second	Supplement Genetic Map	
Partie Parent, res. en Seree	Supplement Ge Type Genetic Map	Narrati naj ele Brene
P         File         Person	Supplement Ge Type Genetic Map S File (e	garatis_nap_eds <u>Boores.</u>
De Paperson and States a Deer     Salard Marsaia	Supplement Ge Type Genetic Map Size (e Organization (	Igenetic map with Brown
Image: State of Contract, St	Supplement Ge Type Genetic Map Size (e Organization (	karning shi <u>Banna</u> karning ang aki <u>Banna</u> arbaning of Gildonia at Davit <u>+</u>

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 GSR Markers: 0 Total SSR Markers: 0 Total Phenotypes: 1344 Total Phenotypes: 1344 Total Environmentals (per sample): 0 Total Environmentals (per site): 0	Covariate Data (Population St 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 Gamples: 2152 Total Genotypes: 12912 Total AFLP Markers: 0 Total SAPD Markers: 0 Total SPM Markers: 0 Total SSR Markers: 6 Total SSR Markers: 0 Total Phonotypes: 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) GP5 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	Dryobalanops lanceolata	Total Sites: 1 Total Samples: 858 Total Genotypes: 13728 Total AFLP Markers: 0 Total IAPD Markers: 0 Total SNP Markers: 0	Environmental Metric Data Environmental Metric Definitio Genotype Data (SSR)

#### treegenesdb.org

### Genetic maps, association or population studies

TGAD Accession Request Form		
This submission process will allow you to upload gener	ic, phenotypic, and/or er	vironmental data into the
TreeGenes database. Results of the association tests	will be accepted as raw F	P-values and all files will be
publication of the associated manuscript.	can be released mined	ability of these decision
Pasalas		TGAD Accession Request Form
Disase fill in the fields being for data pertaining to Bis	us tanta samolas only	
If your study includes multiple species, each set of sp	ecies data will be collecte	Species
		Please fill in the fields below for data pertaining to Plinus taeda samples only If your study includes multiple species, each set of species data will be collected one at a time.
Genetic Data		
Does your study include Genetic data?		GPS Data
Vice		Select the statement that describes how GPS coordinate data was used in your study:
169	<u>ت</u>	GPS coordinates were not collected in this study.
Phenotype Data		Phenotype Data
Does your study include Phenotype?		Select the statement that describes how Phenotype data was used in your study:
Yes	-	Defined Phenotype measurements were performed on each genotyped -
Environmental Data		Environmental Data
Does your study include Environmental data?		Select the statement that describes how Environmental data was used in your study:
		Environmental measurements were performed on each genotyped indiv -
Yes	<u>.</u>	Association Results Data
		Select the statement that best describes your Association results data:
Co	tinue	[Single Marker
		Continue
		and a second second
This is Transferring Accession Hamilton		
The following form in for the submission of periods of	high files and will provide	
the submitter with a Transforms accessor number is the Transforms capabase. This operation is inter-	Full will reference the map . Not for maps associated	
with a manuscript that will be submitted to a peer re	vieweid journal.	
If you would like to submit the mapping files for a mi faren sublished, strategic below this are new	munoriph that has see any	
	and the second se	
genetic map as one the per invage group. Cetaled	instructions on the	
rotes cause and see to reading to see sets can be	NAME OF CALLS	
Palowing successful submission, you will receive a number on the final screen and via e-mail. Your call date in our colleague database to begin this submis	reeGenes accession agout army must be up to mon process.	
Please reference the TreeGenes database and the received here in the manuscript intended for submit	reference number Isson	
For more information about how to compete this for identify also	n Peace planters	
S Prevary Edited Autors Autor parts for Editations	34	
S Advert annuth a land		
Enall	Distance of the local distance of the	
PL Topon	Add Supp	lemental Data
C. Fare Moon Advas Roca Kara Centego C. Greades Halter	For instructions	on the acceptable formats for these fields, <u>click here</u>
Publication Information	Organization Info	
S 18+ Parent of populate strates of annuality in	- gancement and	
S Journal Garages	Author E	ket, Andrew
Provide at 100	Base I	
· Austral Sta	Paper By	eterns of population structure and apportations to broad scale enviro
Genetic May	Supplement Ge	meis Map 💉
2 Via Riprov, August, Street,	Type	
Cogampation (Research of California & Deve	Genetic Map	
Subort Information	<ul> <li>File </li> </ul>	bigeretic_map_ecks Browse.
	Organization i	University of California at Davis
		Submit Information

#### Obtain TGDR accession number!

#### **TreeGenes Data Repository**

Data	Accordian	Paper Title	Enocios	Data Statistics	Data Eilos
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 Sites: 1 Total Senotypes: 480 Total Genotypes: 419520 Total RAPD Markers: 0 Total SNP Markers: 0 Total SSR Markers: 0 Total Phenotypes: 1344 Total Phenotypes: 1344 Total Environmentals (per sample): 0 Total Environmentals (per sample): 0	Covariate Data (Population Stru 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 Gamples: 2152 Total Genotypes: 12912 Total AFLP Markers: 0 Total IRAPD Markers: 0 Total CpSSR Markers: 0 Total CpSSR Markers: 0 Total SSR Markers: 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) GP5 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 Definition Genotype Data (SSR)



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

- <u>http://tripal.info</u>
- 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)

Dorrie Main, Sook Jung, Stephen Ficklin Washington State University • Genome Database for Rosaceae, • Cool Season Food Legumes	Canada Kirstin Bett, Lacey Sanderson Univ of Saskatchewan • KnowPulse	Tripal Gateway Project Tree (& Legume) Databases
<ul> <li>Citrus Genome Database</li> <li>University of Utah NSF ACI-REF Collaborators</li> <li>Project Pls</li> <li>Data Transfer Collaborators</li> <li>Collaborating Databases</li> <li>Data Analysis Collaborators</li> </ul>	Steve Cannon, Ethy Cannor, Andrew Farmer, NCGR         • LegumeInfo, PeanutBase         Image: Contract of the state	<ul> <li>In the set of the set of</li></ul>

### TreeGenes Database: Interfaces



# Web-based framework (Galaxy) promotes genomics analysis

< ⇒ ⊂ ∆ 🔒	Secure https://use	galaxy.org	\$ ·	🗉 🥝 🗾 😳 🗣 🍐 👧 📥	🔝 🐴
<b>=</b> Galaxy		Analyze Data Workflow Shared Data	<ul> <li>Visualization - Help - Login or Register -</li> </ul>	C	Using 0%
Tools	<u>±</u>			History	C
search tools	0	Galaxy is an open source, web-based	platform for data intensive biomedical	search datasets	
Get Data		research. If you are new to Galaxy <u>sta</u>	<u>rt here</u> or consult our <u>help resources</u> . You can	Unnamed history	
Lift-Over		install your own Galaxy by following t	(omnty)		
Collection Operations	<u>s</u>	from the <u>Tool Shed</u> .		(empty)	
Text Manipulation				🔒 This history has been	deleted
Datamash					
Convert Formats			Tweets by @galaxyproject 0	load your own data o	r get data
Filter and Sort				from an external sour	rce
Join, Subtract and Gro	oup		Galaxy Project @galaxyproject		
Fetch Alignments/Sec	quences		Plan your GCC2017 agenda on your mobile		
NGS: QC and manipul	lation		device: gcc2017.sched.com/mobile-site		
NGS: DeepTools			#usegalaxy		
NGS: Mapping					
NGS: RNA Analysis			• Hontpellier		
NGS: SAMtools			The BID Transic Conversity Contenues DECENTY's large particular development of the December 2010 and the December 201		
NGS: BamTools		Public Calava Convorc			
NGS: Picard		Public Galaxy Servers	Nution Ages		
NGS: VCF Manipulatio	on	and still counting	🔹 App Store 🌗 Google Play 🕼 Google May		
NGS: Peak Calling			Medicin Web Ages for Phones Android & Baciliterry West Android & Baciliterry West Android & Baciliterry West Android & Baciliterry Baces (19,117)		
NGS: Variant Analysis	2	0 😌 0 0 0 0	Highering and Standard and Andread Standard and Andread Standard and Andread Standard and Andread Standard And Andread Andread And		
NGS: RNA Structure			C Annual Control Contr		
NGS: Du Novo			La Carlo and Anna Anna		
NGS: Gemini			Embed View on Twitter		
NGS: Assembly					
NGS: Chromosome Co	onformation				
NGS: Mothur	Internals	PENNSTATE S			
Charlies Charling	intervals	IOLINS HOPKINS HEALTH			
Statistics					

# 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





# BDSS: Big Data Smart Socket

- 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



# 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: CartograTree

treegenesdb.org



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

### TreeGenes Database: Interfaces



- 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

#### **Discovery Environment** 0000 🔘 Data 🚣 Upload 🔹 🕞 New Folder 🛛 😂 Refresh 🛛 🖳 Download 🔹 📝 Edit 👻 🧀 Share 🔹 Search by Name Trash + + Name . Navigation Last Modified Size Details >> genotype.txt 2014 Jan 12 09:55:27 51 KB 4 🗊 archive Select a file or folder to view its details / jobs ] glm\_BLUEs.txt 239 KB 2014 Jan 12 09:55:26 iob-39277-tasseldispatcher-1013350u1-by-ipc 0 ] glm\_ftest.txt 2014 Jan 12 09:55:28 323 KB job-39279-tasseldispatcher-1013350u1-by-ipc Im h run\_pipeline.pl... 2014 Jan 12 09:55:29 0 bytes job-39281-tasseldispatcher-1013350u1-by-ipc 1 run pipeline.pl., 2014 Jan 12 09:55:30 4 KB job-39289-tasseldispatcher-1013350u1-by-ipc tasseldispatch... 2014 Jan 12 09:55:29 4 KB job-39291-tasseldispatcher-1013350u1-by-ipc Þ 🚺 job-39292-tasseldispatcher-1013350u1-by-ipc htasseldispatch... 2014 Jan 12 09:55:28 3 KB job-39319-tasseldispatcher-1013350u1-by-ipc Traits.txt 2014 Jan 12 09:55:26 309 bytes job-39347-tasseldispatcher-1013350u1-by-ipc Coge\_data Sswap Image: Community Data Shared With Me 🖻 🚺 Trash

– Connect with Cyverse Views

Download data locally or maintain on cloud-based storage

### CartograTree: Current Development

- Flexible georeferenced tagging
  Approximate

  - Exact
  - Obscured (radius)
- Environmental layers (Geoserver)
  - Soil
  - Fire/DroughtClimate 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



# CartograTree: Advanced Interface



- 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)

