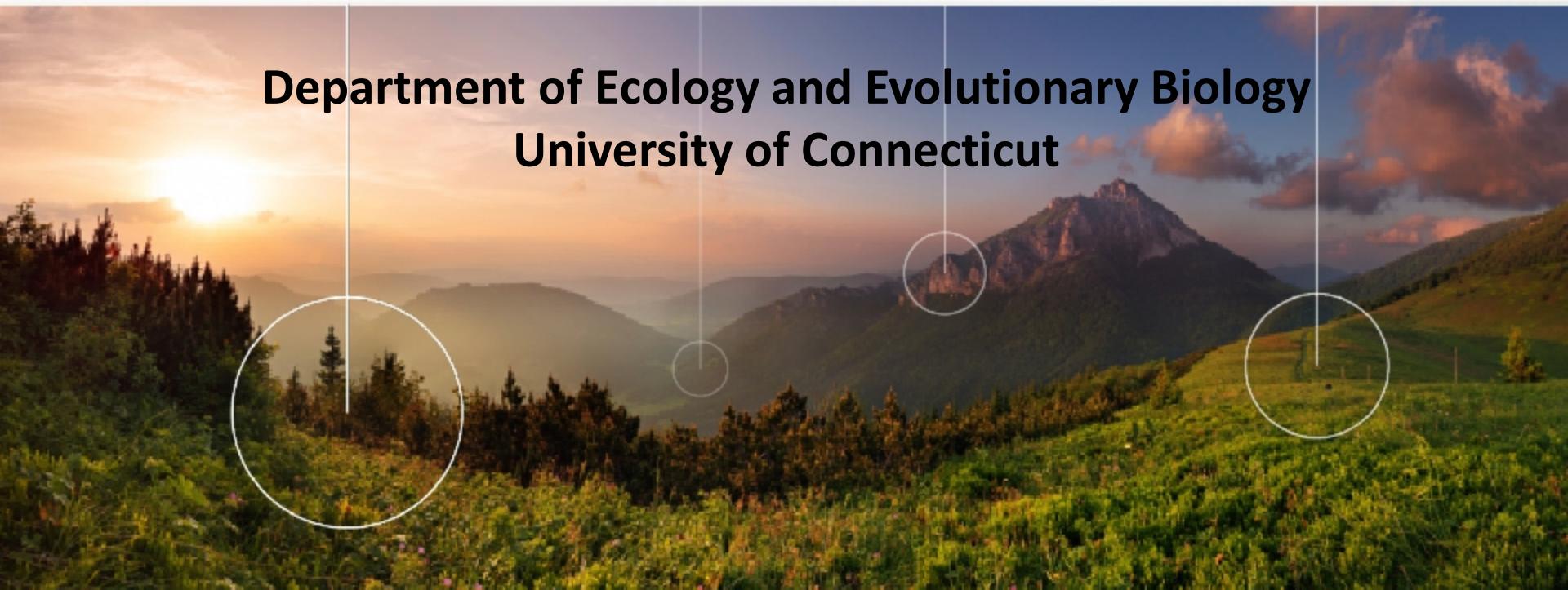




# *Computational approaches to decode megagenomes and develop database resources for the forest tree community*

Jill Wegrzyn



Department of Ecology and Evolutionary Biology  
University of Connecticut

# Data Science: More data or better algorithms?

**Google's Research Director Peter Norvig (2010):**

**"We don't have better algorithms. We just have more data."**



We Want More Data

# Big Data in Genomics

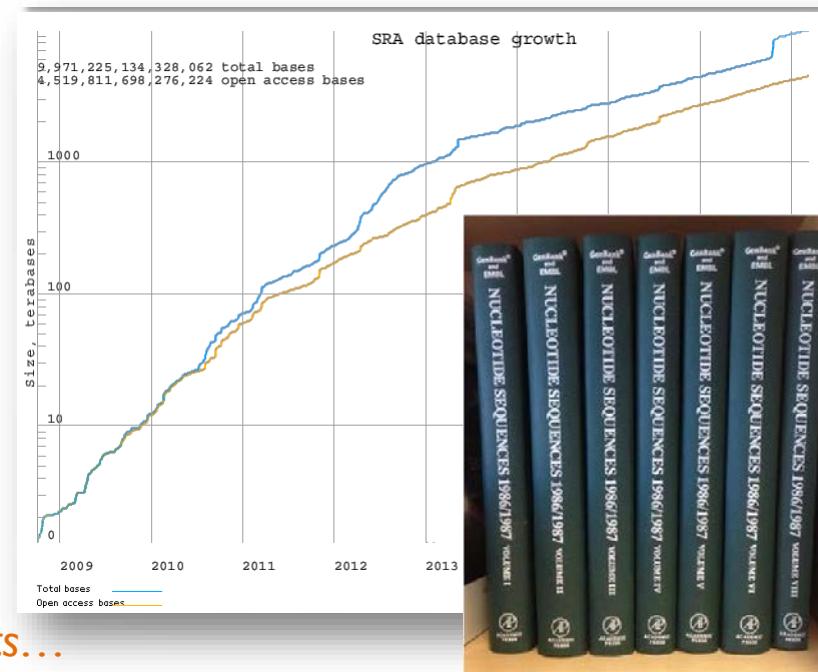
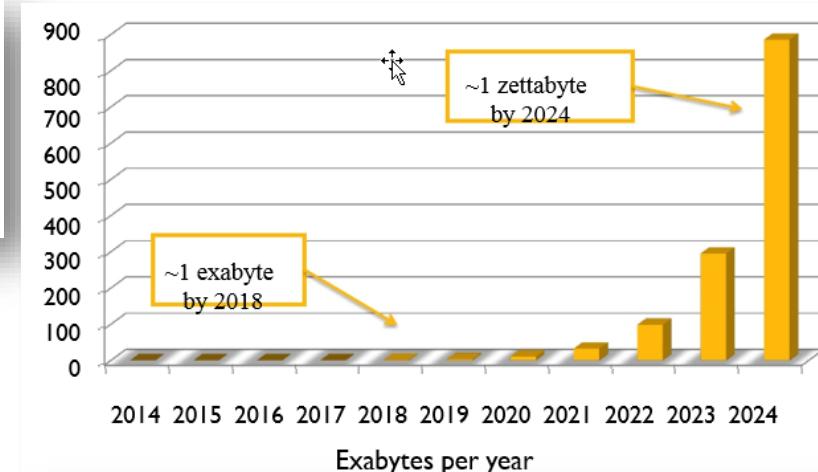
PERSPECTIVE

## 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

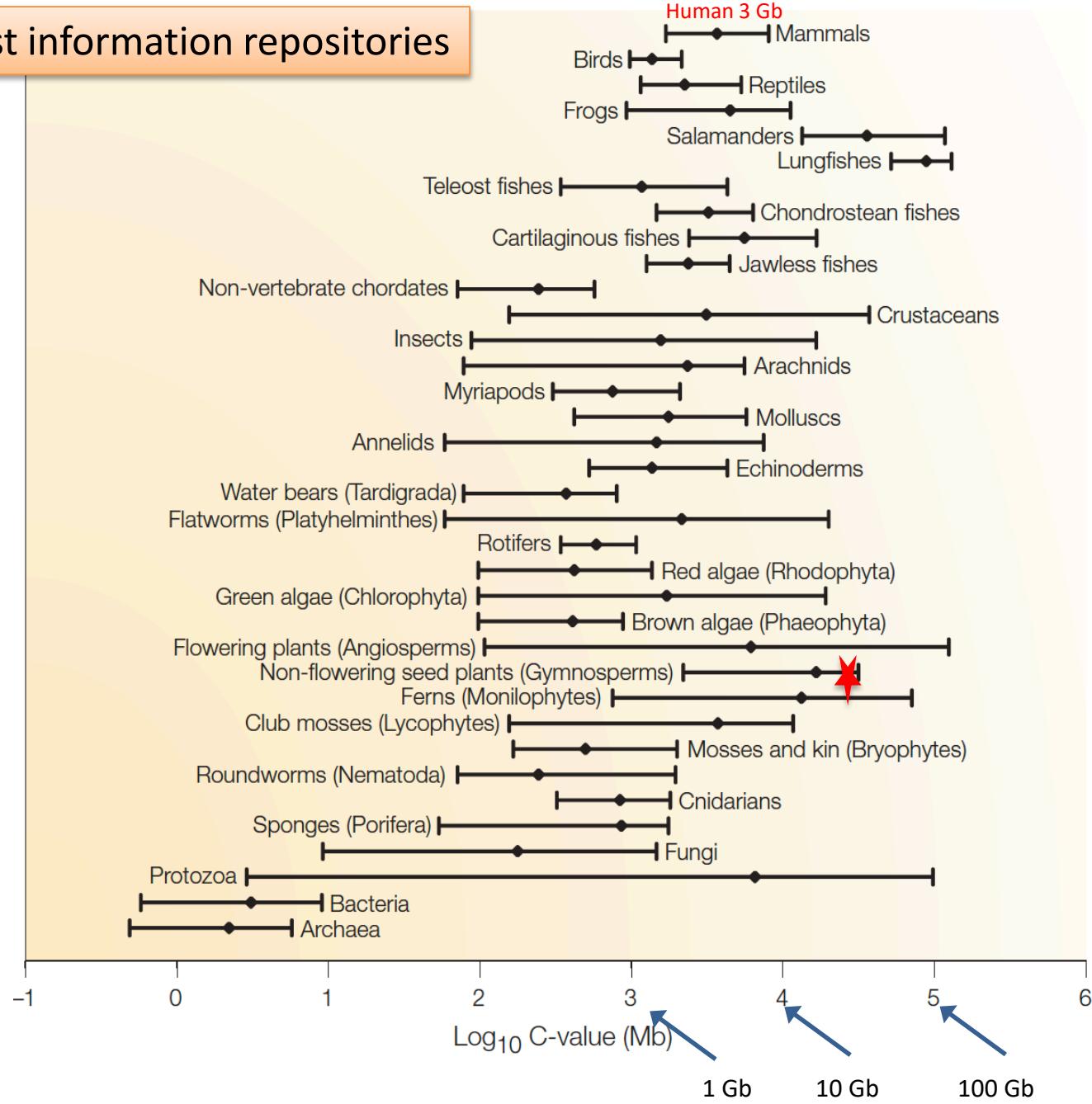
"Compared genomics with three other major generators of Big Data: **Astronomy, YouTube, and Twitter**...Genomics is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis"



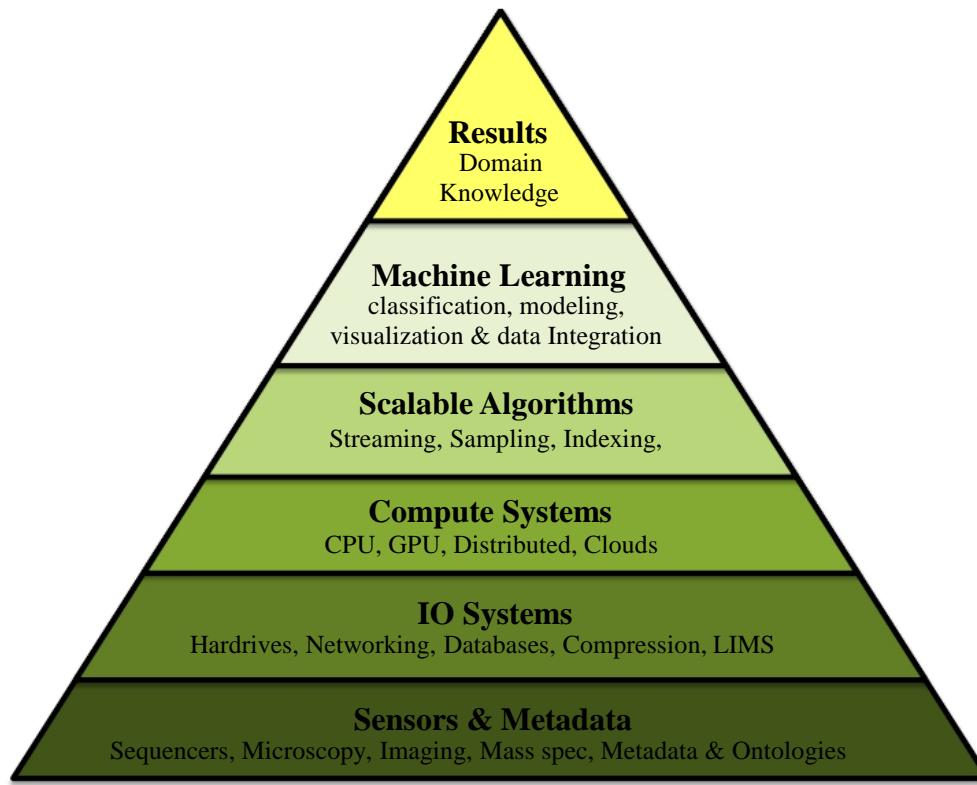
Mostly Genomic but...Proteomics, Phenomics, Metabolomics...

# Genomes are vast information repositories

- Kb = 1000 bp
- Mb =  $1 \times 10^6$  bp
- Gb =  $1 \times 10^9$  bp
- Tb =  $1 \times 10^{12}$  bp
- Pb =  $1 \times 10^{15}$  bp



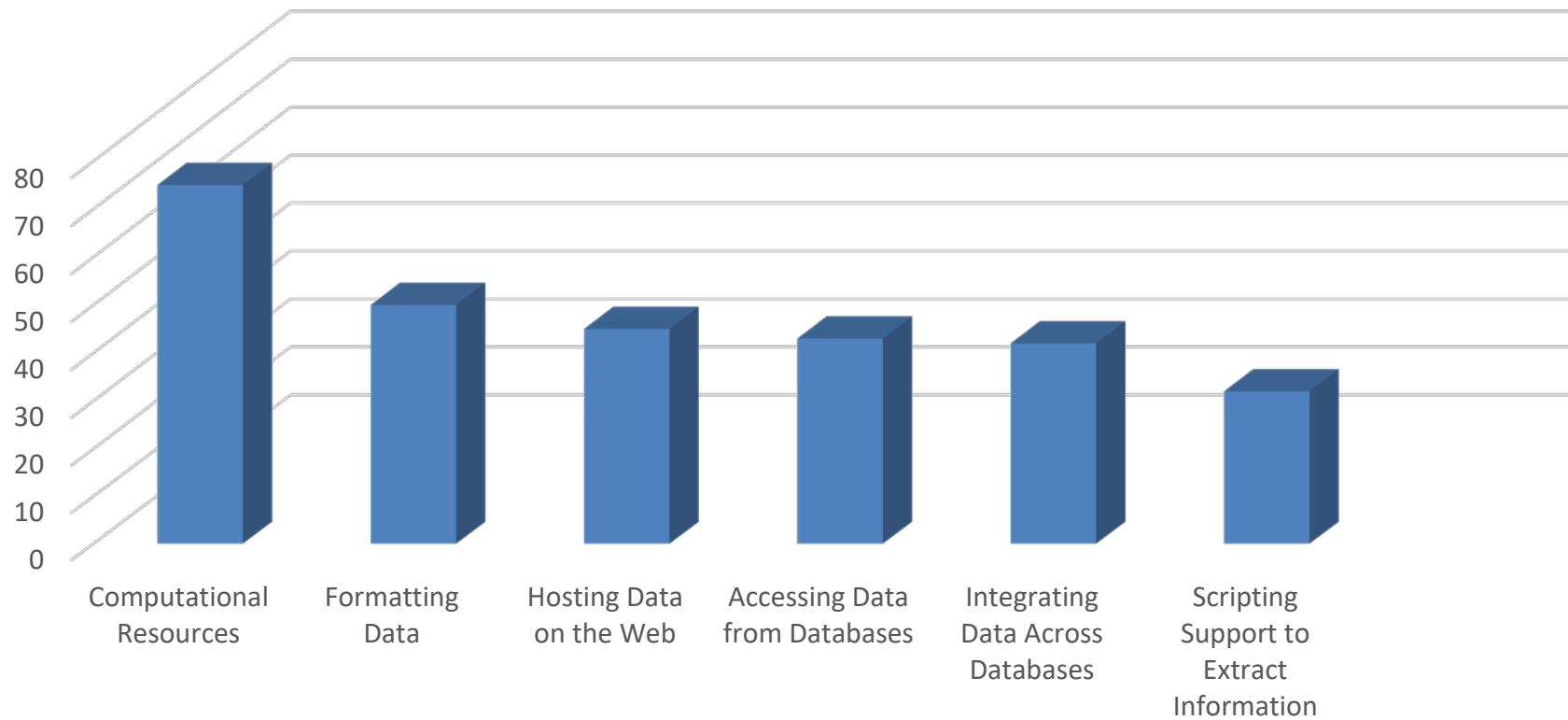
# Acquiring Knowledge through Big Data



# 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 (~1,092 users)

# 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



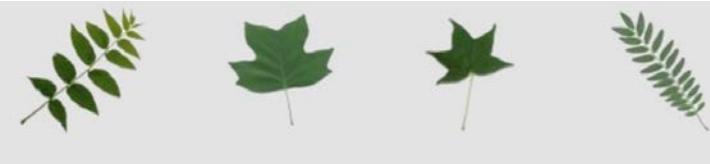
# Tripal

Open source content management system (CMS) for biological data

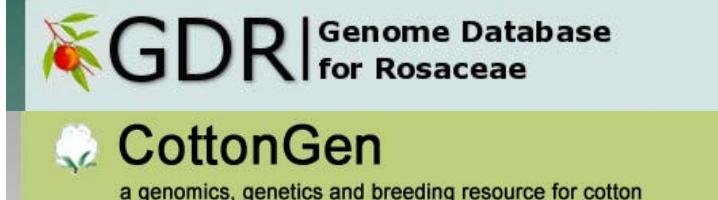
*Modules for genetic, genomic, and breeding data generated through a CMS and standardized schema*

#### Benefits:

- Reduces development costs
- Provides an API for complete customization
- Uses GMOD Chado and community ontologies for standardization
- Allows for sharing of extensions between sites

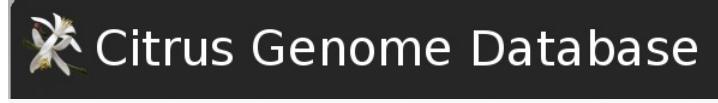


The Hardwood Genomics Project



CottonGen

a genomics, genetics and breeding resource for cotton

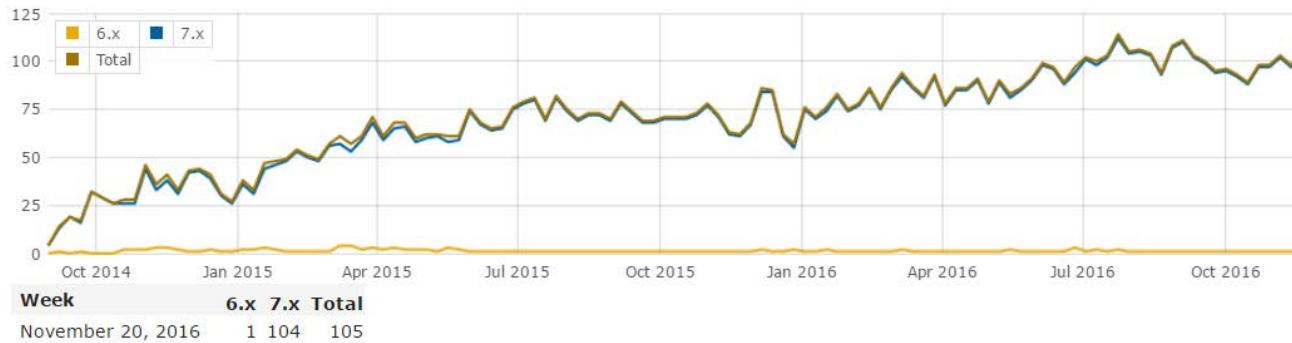


# Current State of Tripal

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



Weekly project usage



# TREEGENES DATABASE



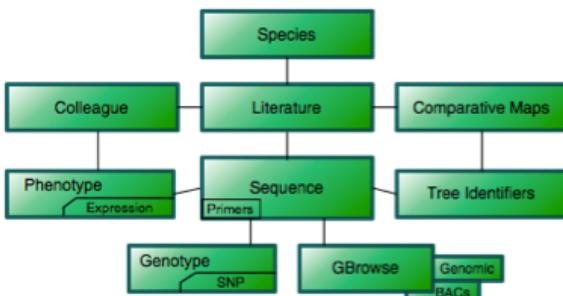
Welcome Research TreeGenes DiversiTee FTGSC iPlant Resources Events News Jobs Links Help

Welcome to the TreeGenes Project!



## TreeGenes :: Overview

The TreeGenes database and Dendrome project provide custom informatics tools to manage the flood of information resulting from high-throughput genomics projects in forest trees from sample collection to downstream analysis. This resource is enhanced with systems that are well connected with federated databases, automated data flows, machine learning analysis, standardized annotations and quality control processes. The database itself contains several curated modules that support the storage of data and provide the foundation for web-based searches and visualization tools. **GMOD** GUI tools such as **CMAP** for genetic maps and **GBrowse** for genome and transcriptome assemblies are implemented here. A sample tracking system, known as the **Forest Tree Genetic Stock Center**, sits at the forefront of most large-scale projects. Barcode identifiers assigned to the trees during sample collection are maintained in the database to identify an individual through DNA extraction, resequencing, genotyping and phenotyping.



## Browser Portal

Currently, TreeGenes uses two web based browsers for sequence and annotation visualization. First, the Generic Genome Browser (GBrowse v. 2.54) is a combination of database and interactive website for manipulating and displaying annotations on genomes.

The second browser, WebApollo (v. 052013), is built upon JBrowse. From WebApollo, a variety of annotation tracks are available for visualization against the draft assembly of the Loblolly Pine (*Pinus taeda*) genome.

## GENOME

### Betula nana

[Downloads](#) [More Info](#)

### Eucalyptus camaldulensis

[Downloads](#) [More Info](#)

### Eucalyptus grandis

[GBrowse](#) [More Info](#)

### Fraxinus excelsior

[Downloads](#) [More Info](#)

### Manihot esculenta

[GBrowse](#) [More Info](#)

### Picea abies

[Downloads](#) [More Info](#)

### Picea glauca

[More Info](#)

### Pinus taeda

[Browsers](#) [More Info](#) [Downloads](#)

[GBrowse - BACs](#) »

[GBrowse - Fosmids](#) »

[WebApollo - Annotated Scaffolds](#) »

Login: demo

Password: demo

# TreeGenes Database: Species

[treegenesdb.org](http://treegenesdb.org)

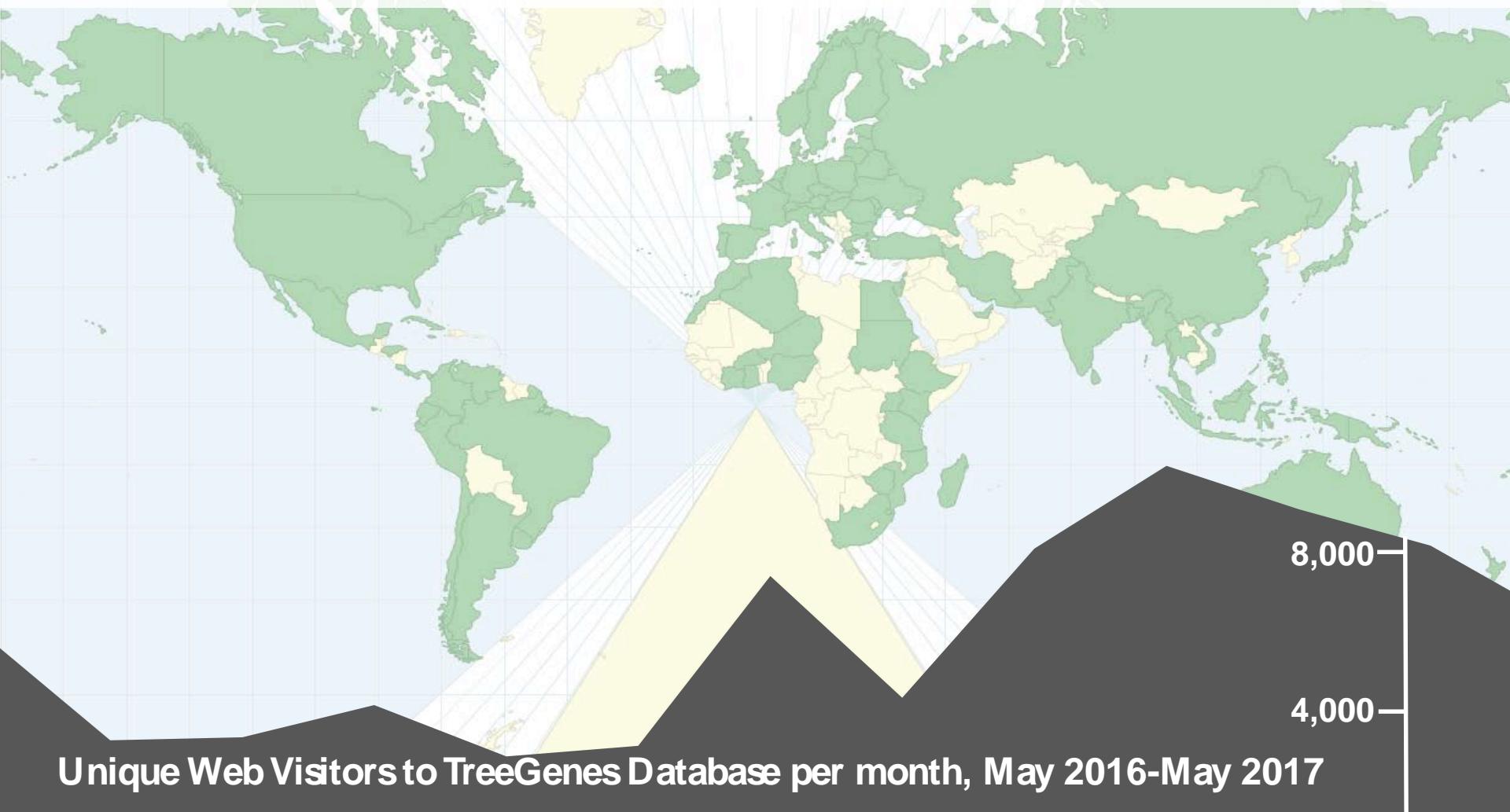


- 1,701 species from 112 genera
  - At least one genetic artifact from each species
  - Conifers but is currently inclusive of all forest trees
- Full genome sequence: 15 species
- Transcriptome/Expression resources: 6,920,817 sequences from 322 species
- 108 genetic maps from 37 species
- Extensive genotypic data (GBS and array)

# TreeGenes Database: Users

[treegenesdb.org](http://treegenesdb.org)

2,012 users from 855 organizations in 92 countries



# New TreeGenes Coming Soon!

 **TreeGenes**  
A Forest Tree Genome Database

My account Log out

Home    FTP Access    About TreeGenes    Contact

**Data**

- ▼ [FTP data download](#)
  - [Genomes](#)
  - [Transcriptomes](#)
  - [Genotyping](#)
  - [TreeGenes UniGene](#)
- [Species](#)
- [Literature](#)

**Tools**

- [CartograTree](#)
- [JBrowse](#)
- [DiversiTTree](#)
- [CMap](#)





Welcome to the new Tripal-powered TreeGenes site!  
We're working on getting everything ready for you.

**About TreeGenes**

The TreeGenes database provides custom informatics tools to manage the flood of information resulting from high-throughput genomics projects in forest trees from sample collection to downstream analysis. This resource is enhanced with systems that are well connected with federated databases, automated data flows, machine learning analysis, standardized annotations and quality control processes. The database itself contains several curated modules that support the storage of data and provide the foundation for web-based searches and visualization tools.

[Read more](#)

**TreeGenes on Twitter**

**Tweets** by @TreeGenes

 TreeGenes Retweeted 

 EMBL-EBI  
@emblebi

Wheat genome data now available in Ensembl Plants!  
[plants.ensembl.org/Triticum\\_aestivum](#)...

  19 Apr

 TreeGenes Retweeted 

 AgBioData  
@AgBioData

The AgBioData website is live!  
[agbiodata.org](#)

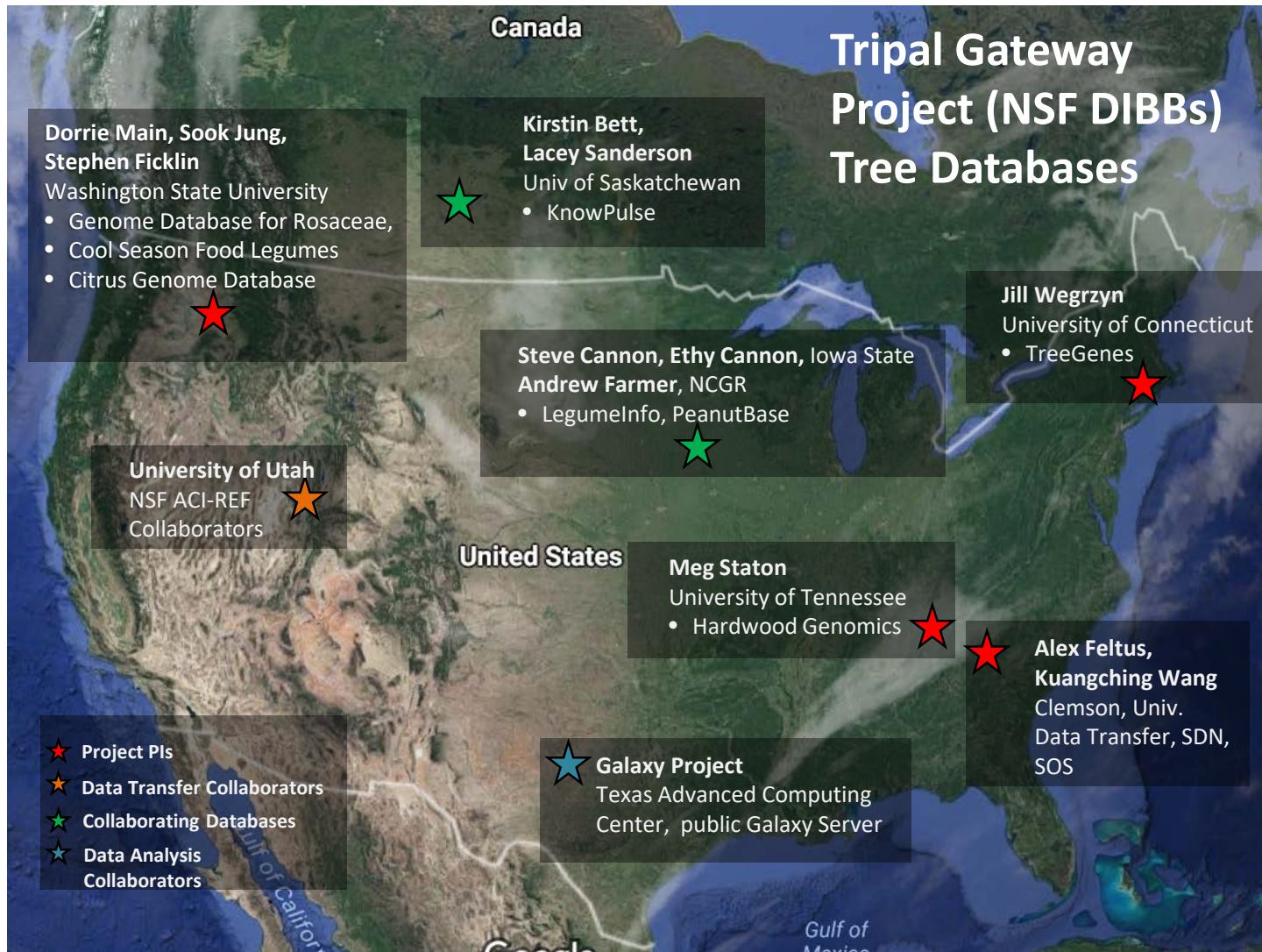
 AgBioData  
A global network for genomics, genetics and breeding research in agriculture and the environment

 AgBioData Workshop  
An international workshop for plant breeders and genomicists to share their knowledge and expertise, and to facilitate the development of new plant breeding methods and applications.

# 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)

# Tripal Gateway Project (NSF DIBBs) Tree Databases



# What is Galaxy?

Secure | <https://usegalaxy.org>

Analyze Data Workflow Shared Data Visualization Help Login or Register Using 0%

**Galaxy**

Tools

- search tools
- [Get Data](#)
- [Lift-Over](#)
- [Collection Operations](#)
- [Text Manipulation](#)
- [Datamash](#)
- [Convert Formats](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Fetch Alignments/Sequences](#)
- [NGS: QC and manipulation](#)
- [NGS: DeepTools](#)
- [NGS: Mapping](#)
- [NGS: RNA Analysis](#)
- [NGS: SAMtools](#)
- [NGS: BamTools](#)
- [NGS: Picard](#)
- [NGS: VCF Manipulation](#)
- [NGS: Peak Calling](#)
- [NGS: Variant Analysis](#)
- [NGS: RNA Structure](#)
- [NGS: Du Novo](#)
- [NGS: Gemini](#)
- [NGS: Assembly](#)
- [NGS: Chromosome Conformation](#)
- [NGS: Mothur](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#). You can install your own Galaxy by following the [tutorial](#) and choose from thousands of tools from the [Tool Shed](#).

080 +

Public Galaxy Servers and *still* counting

Embed View on Twitter

History

- search datasets
- Unnamed history (empty)
- This history has been deleted
- This history is empty. You can [load your own data](#) or [get data from an external source](#)

PENNSTATE JOHN HOPKINS UNIVERSITY OREGON HEALTH & SCIENCE UNIVERSITY TACC CYVERSE

# Galaxy Integration



- Galaxy-Tripal crosstalk: Blend4php
  - PHP library, independent of Tripal that provides a wrapper for the Galaxy API
  - Any PHP application can interact with Galaxy
  - <https://github.com/galaxyproject/blend4php>
  - Provides a full suite of unit tests!

A screenshot of a GitHub repository page for 'galaxyproject / blend4php'. The page shows basic repository statistics: 444 commits, 5 branches, 1 release, and 4 contributors. It also displays a summary message about the PHP API and a link to the documentation. At the bottom, there are buttons for creating new files, uploading files, finding files, and cloning or downloading the repository. A commit message from 'spficklin' is visible at the bottom left, stating 'Unit tests completing successfully for alpha release'. The GitHub interface includes standard navigation elements like 'Code', 'Issues', 'Pull requests', 'Pulse', and 'Graphs'.

# Integrating Galaxy with Tripal

Screenshot of a web browser showing the "Manage Galaxy Instances" page in a Tripal overlay.

The URL in the address bar is `192.168.233.165/node#overlay=admin/tripal/extension/galaxy/add`.

The page title is "Manage Galaxy Instances".

Form fields include:

- Galaxy Server Name \***: Input field with placeholder "Please provide the name of the remote Galaxy Server".
- Description**: Text area for additional details about the server.
- URL \***: Input field with placeholder "The URL for the remote Galaxy server".
- User Name**: Input field with placeholder "The user name for the Galaxy server. This username is used to launch all jobs by default. If this field is left blank then it is expected that the user has an account on the Galaxy server and will provide their username when executing workflows."
- API Key**: Input field with placeholder "The API key for the user name specified above. If this field is left blank then it is expected that the user will provide their own API key when submitting a job."

Logos for **amazon web services** and **jetstream** are displayed on the right side of the page.

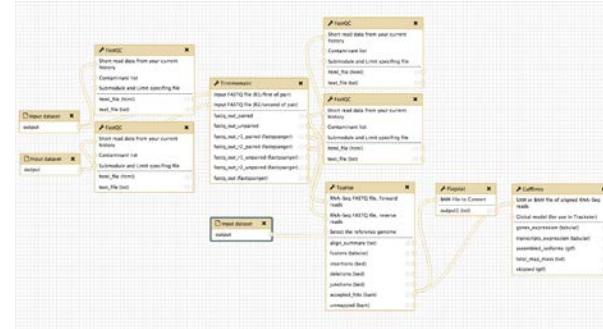
# Galaxy Workflows



Testing on Galaxy instances at Washington State University (GDR), University of Connecticut (TreeGenes), and University of Tennessee (HWG)

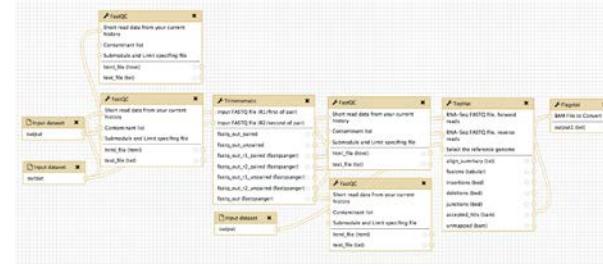
## DNA Sequence Data

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



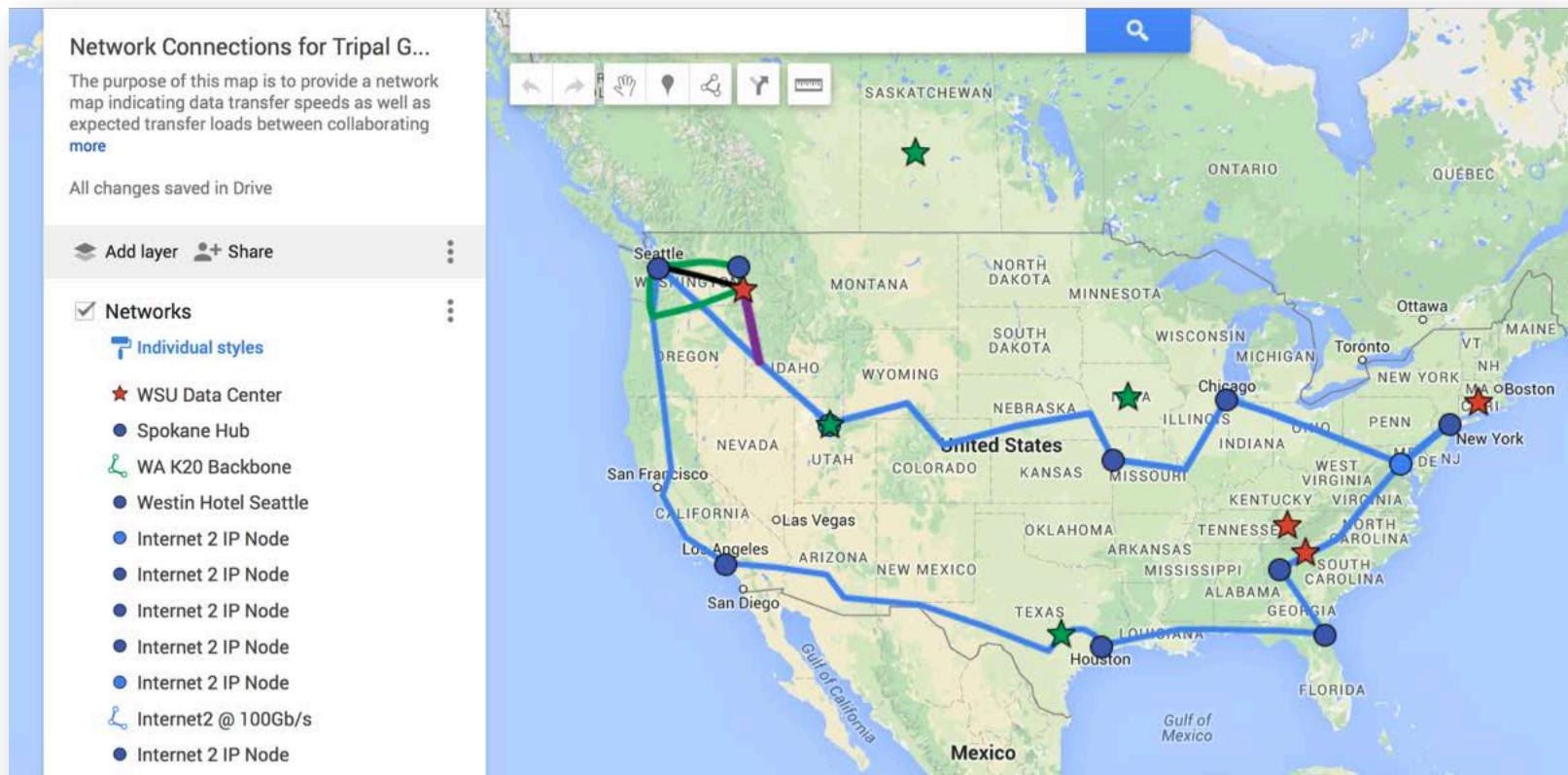
## RNA Sequence Data

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



# TreeGenes Database: Software Defined Networking

[treegenesdb.org](http://treegenesdb.org)



# Big Data Smart Socket

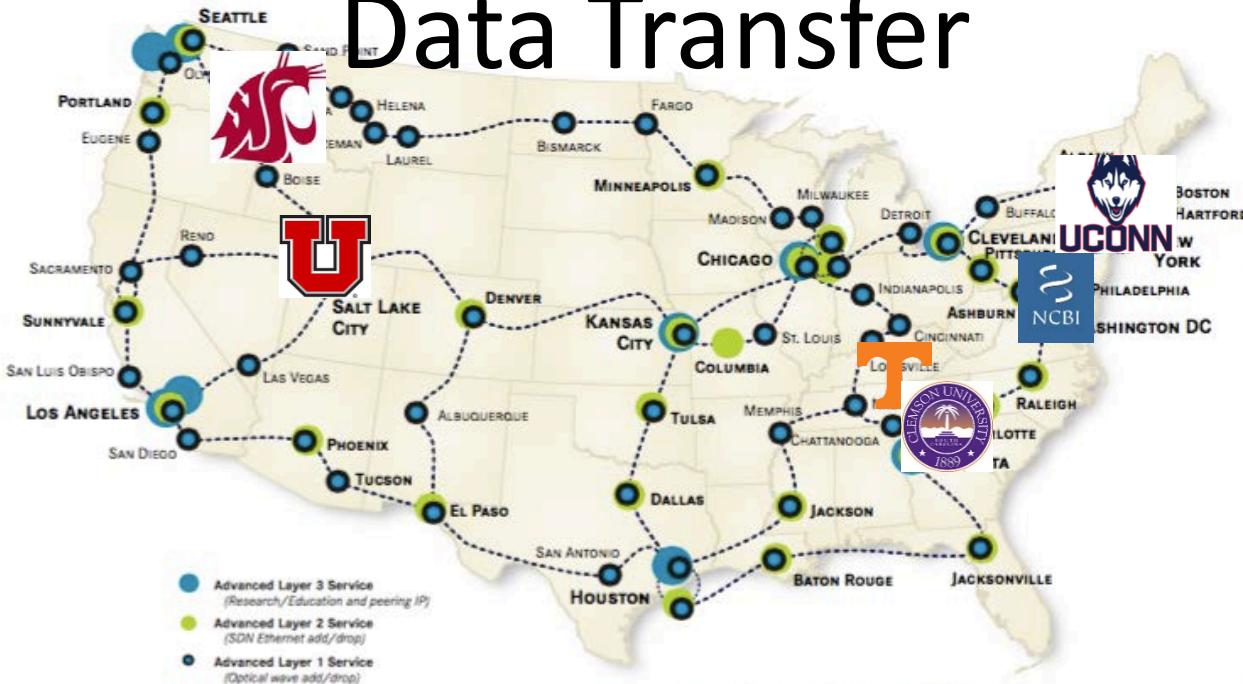
- Smart Data Transfer
- Standalone client with a metadata repository
- First step is to build an inventory of data sources relevant to a particular user community
  - NCBI (Genbank for Raw Data)
  - Cyverse (iPlant for analytics)
  - Tripal supported websites for supporting data
- Determines optimal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



## INTERNET2 NETWORK INFRASTRUCTURE TOPOLOGY

OCTOBER 2014

# Data Transfer



### INTERNET2 NETWORK BY THE NUMBERS

17	JUNIPER MX960 ROUTERS SUPPORTING LAYER 3 SERVICE
34	BROADCOM AND JUNIPER SWITCHES SUPPORTING LAYER 2 SERVICE
62	CUSTOM COLLOCATION FACILITIES
250+	AMPLIFICATION RACKS
15,717	MILES OF NEWLY ACQUIRED DARK FIBER
8.8	TERRA OF OPTICAL CAPACITY
100	GIGS OF HYBRID LAYER 2 AND LAYER 3 CAPACITY
300+	Ciena ActiveFlex 5500 NETWORK ELEMENTS
2,400	MILES PARTNERED CAPACITY WITH ZAYD COMMUNICATIONS IN SUPPORT OF THE NORTHERN TIER REGION



IN SUPPORT OF  
**U.S. UCAN**

NETWORK PARTNERS

ciena

CISCO

INDIANA UNIVERSITY

Infinera

JUNIPER  
NETworks



# Tripal Gateway Use Cases

Researchers often focus on a single gene family and how it evolves across phylogenetic lineages.

Tripal Gateway:

1. A user could search across community DBs for their gene of interest (by BLAST or by functional annotation keyword) using [Tripal Exchange](#).
2. The 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 gene list into the [Tripal Galaxy](#) module.
4. Basic workflow with multiple sequence alignment and phylogenetic tree building could be selected.

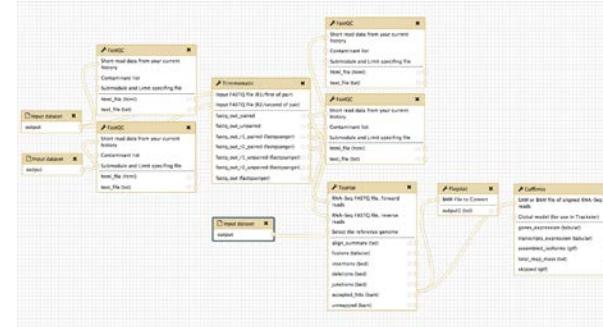
# Galaxy Workflows



Testing on Galaxy instances at Washington State University (GDR), University of Connecticut (TreeGenes), and University of Tennessee (HWG)

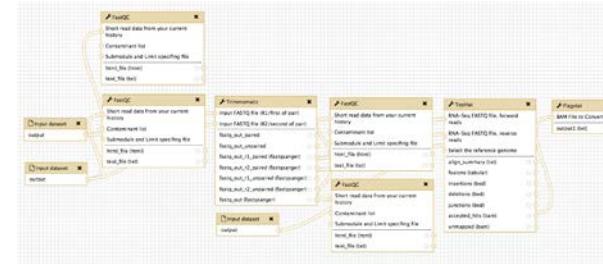
## DNA Sequence Data

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



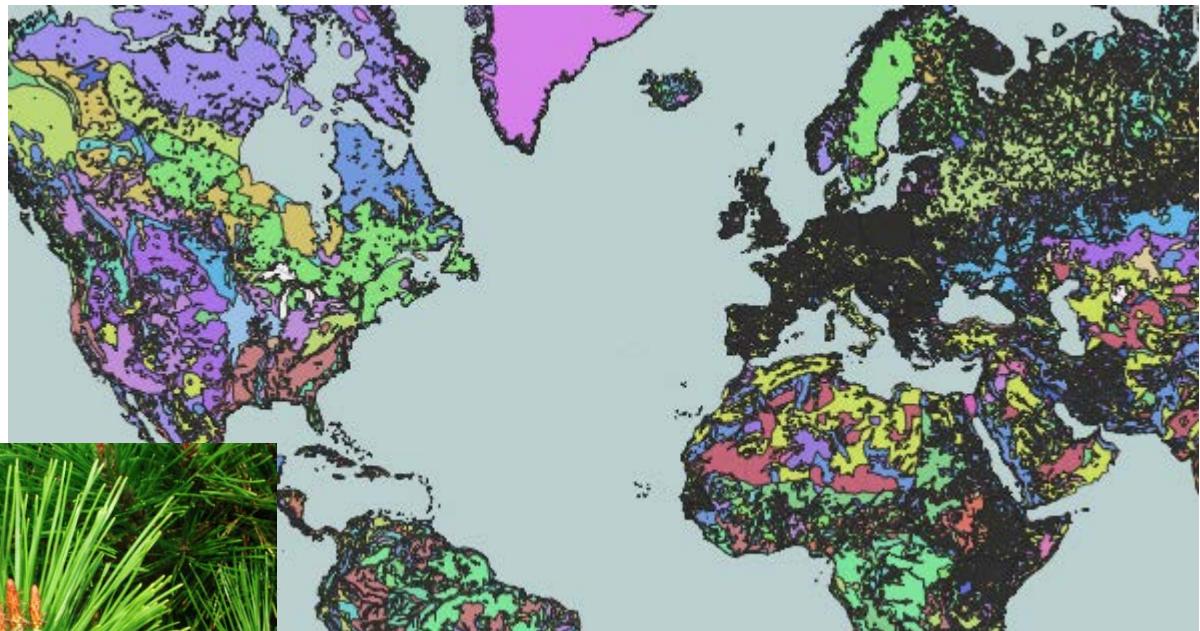
## RNA Sequence Data

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



# Association mapping

TCCTGGAAATGCGATG  
TCACCCATGAATGCGATG  
TGAAAACAAGATGCATG  
GCTGCTGCTCTCCGGGGGG  
GCCCTGGAGGGTGGCCCC  
GCATATGCAGGAAGCGG  
GCCTCCTGACTTTCTCC  
CTCCCAGGCCAGTCCC  
AGCTCGGGAGG





# Drought and pests/pathogens changing the landscape





$$\text{Phenotype} = \text{Genotype} + \text{Environment}$$

---

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



About TreeGenes DiversiTTree Example Contact Credits

**Map Display**

Search for a tree id

Ctrl+Click or Cmd+Click for multiple selections

**All**

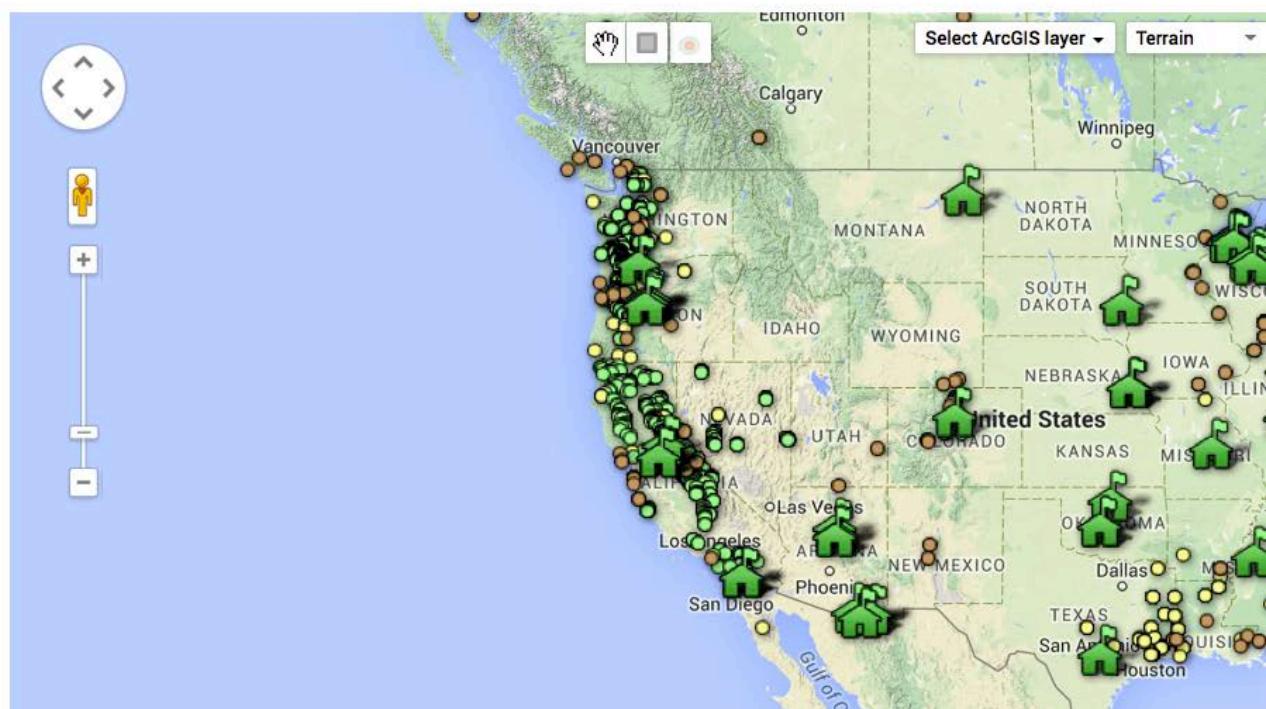
- ▶ Published Studies
- ▶ Taxa
- ▶ Environmental
- ▶ Phenotypes

**Filter Map Display**

Sequenced (83) ?

Genotyped (2940) ?

Phenotyped (9385) ?



- Providing context to geo-referenced data
- Originated from Tree Biology Working Group through iPlant

# TreeGenes Database: CartograTree

treegenesdb.org



About TreeGenes DiversiTree Example Contact Credits

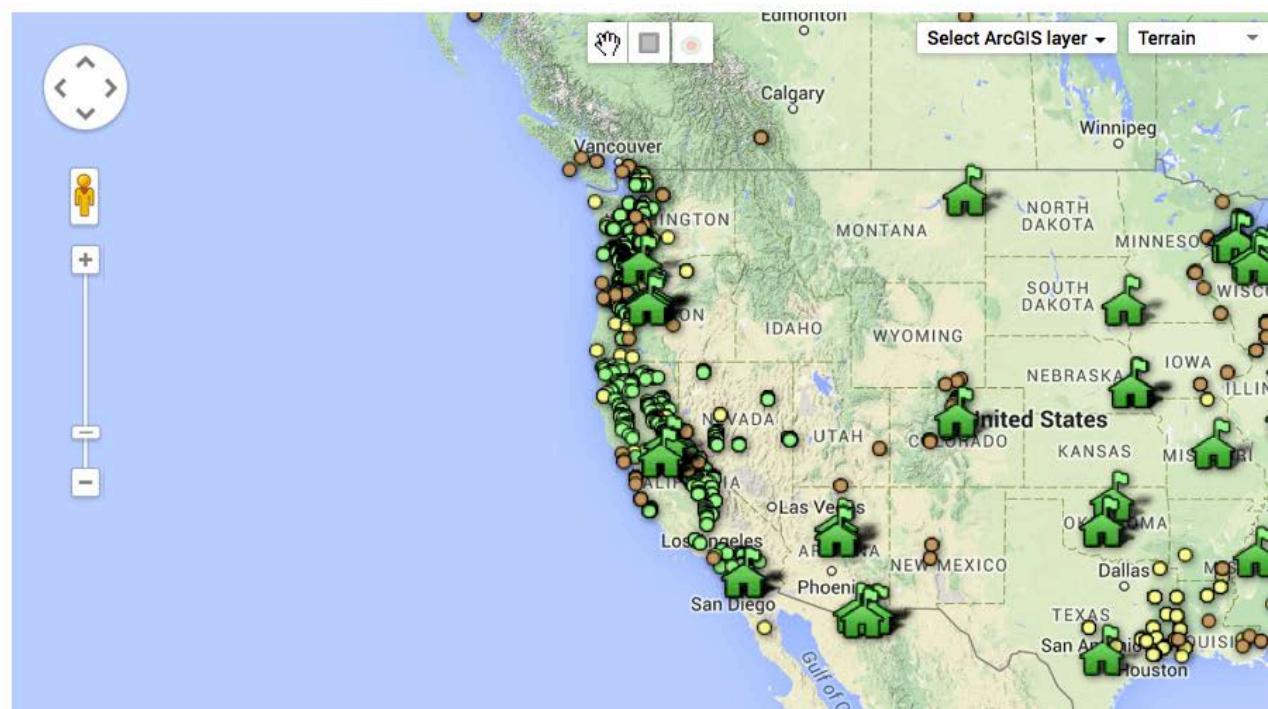
**Map Display**

Search for a tree id

Ctrl+Click or Cmd+Click for multiple selections

**All**

- ▶ Published Studies
- ▶ Taxa
- ▶ Environmental
- ▶ Phenotypes



**Filter Map Display**

Sequenced (83) ?  
 Genotyped (2940) ?  
 Phenotyped (9385) ?

- Data from TreeGenes, WorldClim, Ameriflux, TRY-db
- Google fusion tables & Google maps

# TreeGenes Database: Interfaces

treegenesdb.org

The screenshot shows the TreeGenes Database interface. At the top left, there are two filter options: "Exact (15)" and "Approximate (68)". Below the filters is a Google map of the United States, specifically highlighting the western and southern regions. The map shows several green dots representing gene locations, with labels for major cities like Los Angeles, Phoenix, and San Diego. A legend indicates that green dots represent exact matches and yellow dots represent approximate matches. The map also includes state boundaries and a scale bar of 500 km. Below the map, a section titled "Analyze the data" contains a table with the following columns: ID, Amplicon ID, Top Blast Description (BLAST nr), Species-Specific BLASTs, GO Annotations, Interpro Annotations, and PFAM Annotation. The table lists six rows of data, each with a checked checkbox in the first column. The last column of the table has a tooltip that says "Discover pipelines at SSWAP".

ID	Amplicon ID	Top Blast Description (BLAST nr)	Species-Specific BLASTs	GO Annotations	Interpro Annotations	PFAM Annotation
1	0_10054_01	"NAC domain protein: IPR003441 [Po... "NAC domain protein: IPR00...	"NAC domain protein: IPR00...	"GO:0003676 nucl... -999		-999
2	0_10706_01	"uninformative (1.90E-27) (unspecified)" "uninformative (1.90E-27) (u...	"uninformative (1.90E-27) (unspecified)" "uninformative (1.90E-27) (u...	"GO:0009536 plas... -999		-999
3	0_11090_01	"PREDICTED: protein OBERON 4-like..." "PREDICTED: protein OBER...	"PREDICTED: protein OBERON 4-like..." "PREDICTED: protein OBER...	-999	"IPR004082 Protein of u...	-999
4	0_11270_01	"PREDICTED: probably inactive leuci..." "PREDICTED: probably inact...	"PREDICTED: probably inactive leuci..." "PREDICTED: probably inact...	"GO:0016740 tran... -999		-999
5	0_11389_01	"PREDICTED: protein phosphatase 2..." "PREDICTED: protein phosph...	"PREDICTED: protein phosphatase 2..." "PREDICTED: protein phosph...	"GO:0001932 reg... -999		-999
6	0_11411_01	"PREDICTED: galactose oxidase-like..." "PREDICTED: galactose oxi...	"PREDICTED: galactose oxidase-like..." "PREDICTED: galactose oxi...	-999	"IPR013783 Immunologic...	"PF00118 Domain o...

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

# TreeGenes Database: SSWAP

treegenesdb.org



New pipeline - executed

Logged as hansvg  
[My pipelines](#) [Logout](#)



Output Data Set



Display Data: [Click here to send the output data to a viewer \(opens in a new window\)](#)  
iPlant Path: [/hansvg/sswap/New\\_pipeline-2014-01-12-10-38-42.owl](/hansvg/sswap/New_pipeline-2014-01-12-10-38-42.owl)  
Produced by: <http://sswap.info/iplant/svc/tassel>

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

# TreeGenes Database: Cyverse (TACC)

The screenshot shows the 'Data' view of the TreeGenes Database on the Cyverse platform. The interface includes a top navigation bar with 'Upload', 'New Folder', 'Refresh', 'Download', 'Edit', 'Share', and a 'Search by Name' field. Below this is a 'Navigation' sidebar with a tree view of available data:

- archive
  - jobs
    - job-39277-tasseldispatcher-1013350u1-by-ipc
    - job-39279-tasseldispatcher-1013350u1-by-ipc
    - job-39281-tasseldispatcher-1013350u1-by-ipc
    - job-39289-tasseldispatcher-1013350u1-by-ipc
    - job-39291-tasseldispatcher-1013350u1-by-ipc
    - job-39292-tasseldispatcher-1013350u1-by-ipc
    - job-39319-tasseldispatcher-1013350u1-by-ipc
    - job-39347-tasseldispatcher-1013350u1-by-ipc
  - coge\_data
  - sswap
  - tempDir
  - Community Data
  - Shared With Me
  - Trash

The main panel displays a table of files with columns for Name, Last Modified, and Size. The files listed are:

Name	Last Modified	Size
genotype.txt	2014 Jan 12 09:55:27	51 KB
glm_BLUEs.txt	2014 Jan 12 09:55:26	239 KB
glm_ftest.txt	2014 Jan 12 09:55:28	323 KB
run_pipeline.pl...	2014 Jan 12 09:55:29	0 bytes
run_pipeline.pl...	2014 Jan 12 09:55:30	4 KB
tasseldispatch...	2014 Jan 12 09:55:29	4 KB
tasseldispatch...	2014 Jan 12 09:55:28	3 KB
traits.txt	2014 Jan 12 09:55:26	309 bytes

A tooltip message 'Select a file or folder to view its details' is visible in the 'Details' column.

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

# Metadata Needed! Data Integration

## TreeGenes Data Repository



### Sequence Resources

[Summary by Genus](#)

### Colleague Directory

[Colleagues](#)

[Organizations](#)

### Species Database

[Forest Trees](#)

### Literature Database

[Search Literature](#)

### Transcriptome Database

[Search Transcriptome](#)

[Transcriptome Summary](#)

### Protein Database

[Search Proteins](#)

[Protein Summary](#)

### Expression 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	<a href="#">Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (<i>Populus trichocarpa</i>, Salicaceae) secondary xylem.</a>	<i>Populus trichocarpa</i>	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	<a href="#">Covariate Data (Population Structure)</a> <a href="#">Genotype Data (SNP)</a> <a href="#">GPS Data</a> <a href="#">Haplotype Data</a> <a href="#">Phenotype Data</a> <a href="#">Phenotype Definitions</a>
9/25/2012	TGDR002	<a href="#">Astonishingly low genetic variation in <i>Quercus acutissima</i>, an important tree species in Satoyama, a traditional Japanese rural forest and agricultural landscape, revealed by chloroplast microsatellite markers</a>	<i>Quercus acutissima</i>	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	<a href="#">Genotype Data (cpSSR)</a> <a href="#">GPS Data</a> <a href="#">Haplotype Data</a> <a href="#">Supplemental Data</a>
11/5/2012	TGDR003	<a href="#">Extensive selfing in an endangered population of <i>Pinus parviflora</i> var. <i>parviflora</i> (Pinaceae) in the Boso Hills, Japan</a>	<i>Pinus parviflora</i>	Total Sites: 2 Total Samples: 116 Total Genotypes: 464 Total AFLP Markers: 0 Total RAPD Markers: 0 Total SNP Markers: 0	<a href="#">Genotype Data (SSR)</a> <a href="#">GPS Data</a> <a href="#">Supplemental Data</a> <a href="#">Supplemental Data</a> <a href="#">Supplemental Data</a> <a href="#">Supplemental Data</a>

# Association mapping with CartograTree



**HWG**

Hardwood Genomics Project

ArcGIS

Harmonized World Soil Dataset - Major Soil Groups  
(Data Basin Dataset)

## GENOME DATABASE FOR ROSACEAE



Resources for Rosaceae Research Discovery and Crop Improvement

Sign In

**WorldClim - Global Climate Data**  
*Free climate data for ecological modeling and GIS*



# Association mapping with CartograTree

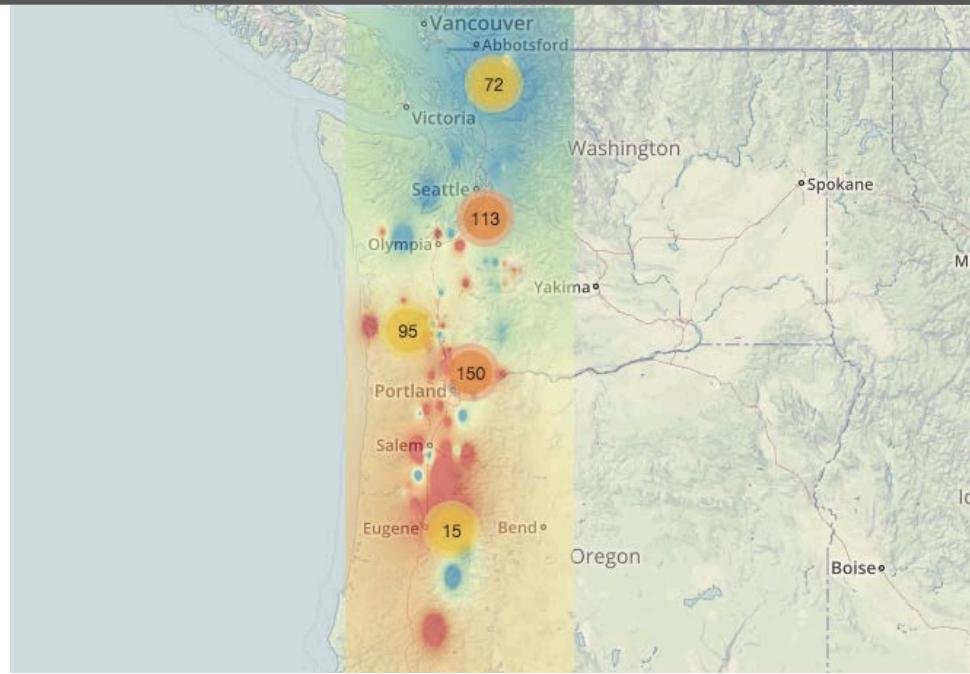


# TreeGenes Database: Interfaces

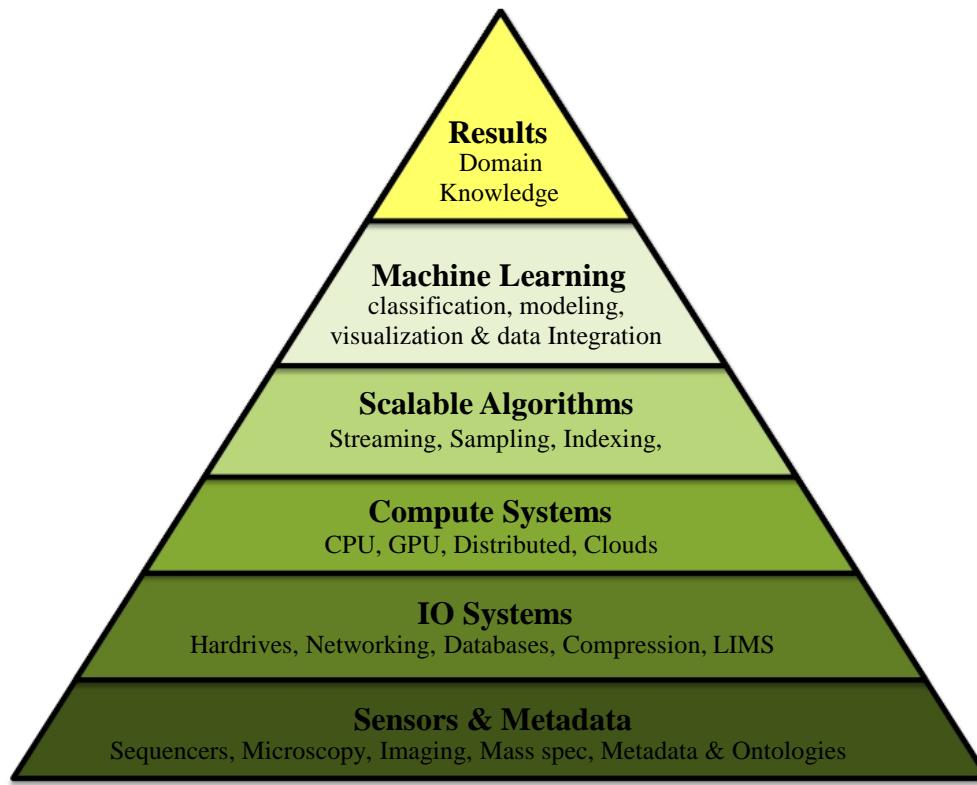
[treegenesdb.org](http://treegenesdb.org)

## Current Development

- ***Better integration of layers (soil, climate prediction layers)***
- ***Real time association of genotype to environment***
- Observe gradients and population overlays
- TGDR Data Submission and Galaxy API in Tripal



# Acquiring Knowledge through Big Data

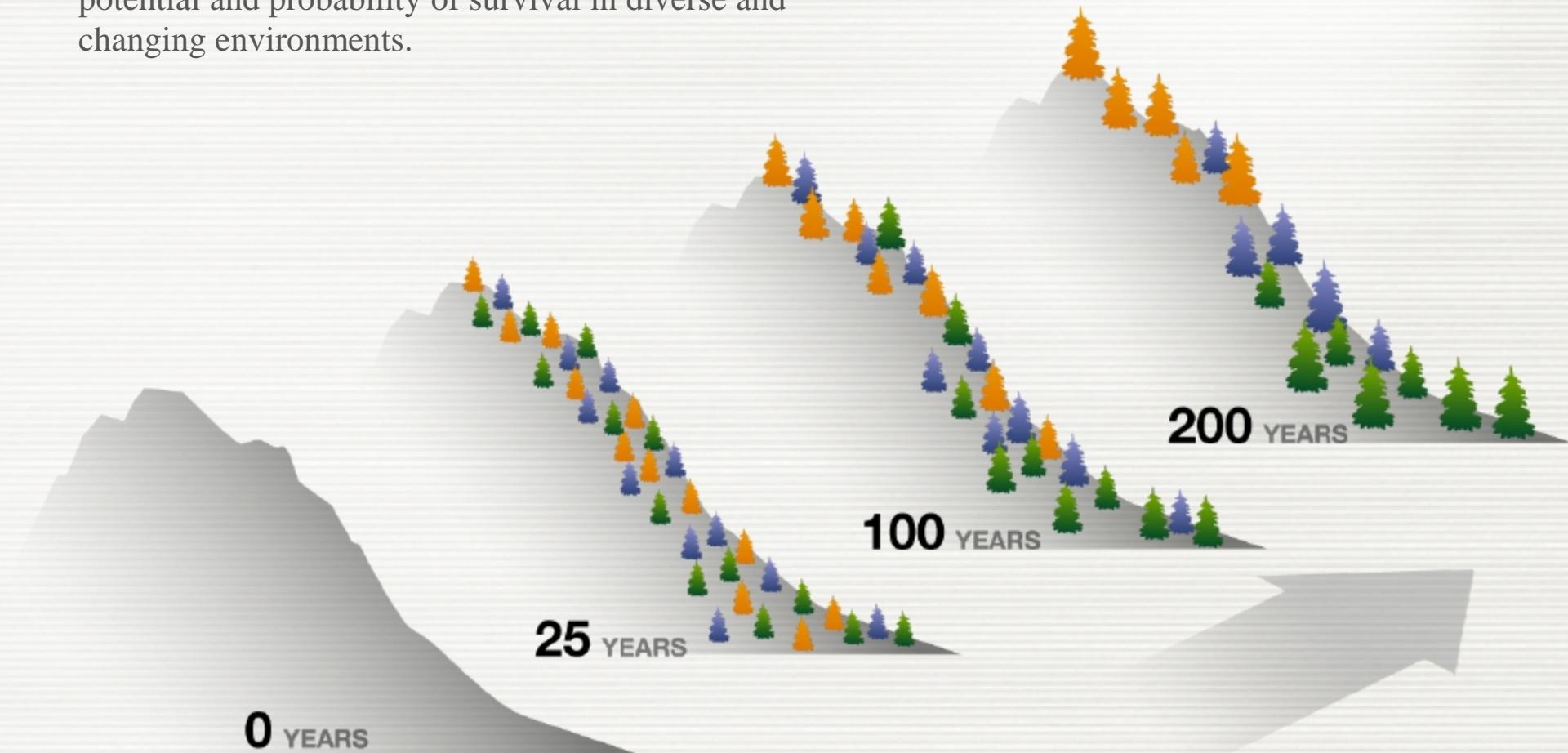




# Adaptive Potential

An organism's genetic makeup determines its adaptive potential and probability of survival in diverse and changing environments.

-  cold tolerant
-  not cold tolerant



# Transcriptomes in Forest Trees

## Evo-devo Study

Dimorphism between juvenile and adult leaves (heteroblasty)  
Juniper (left) and Pine (right).

**UNAM - Lobo**



## Landscape Genomics

Identifying genes and alleles responsible for adaptation along an elevational gradient in two different species (Limber Pine – left, Engelmann Spruce – right).

**Colorado State - Mitton**



## Association Genetics

The *Trojan fir* (Christmas tree) transcriptome is being investigated for disease resistance genes against phytophthora by examining transcriptomes of susceptible and partially resistant trees.

**NCSU – Whetten/Frampton**



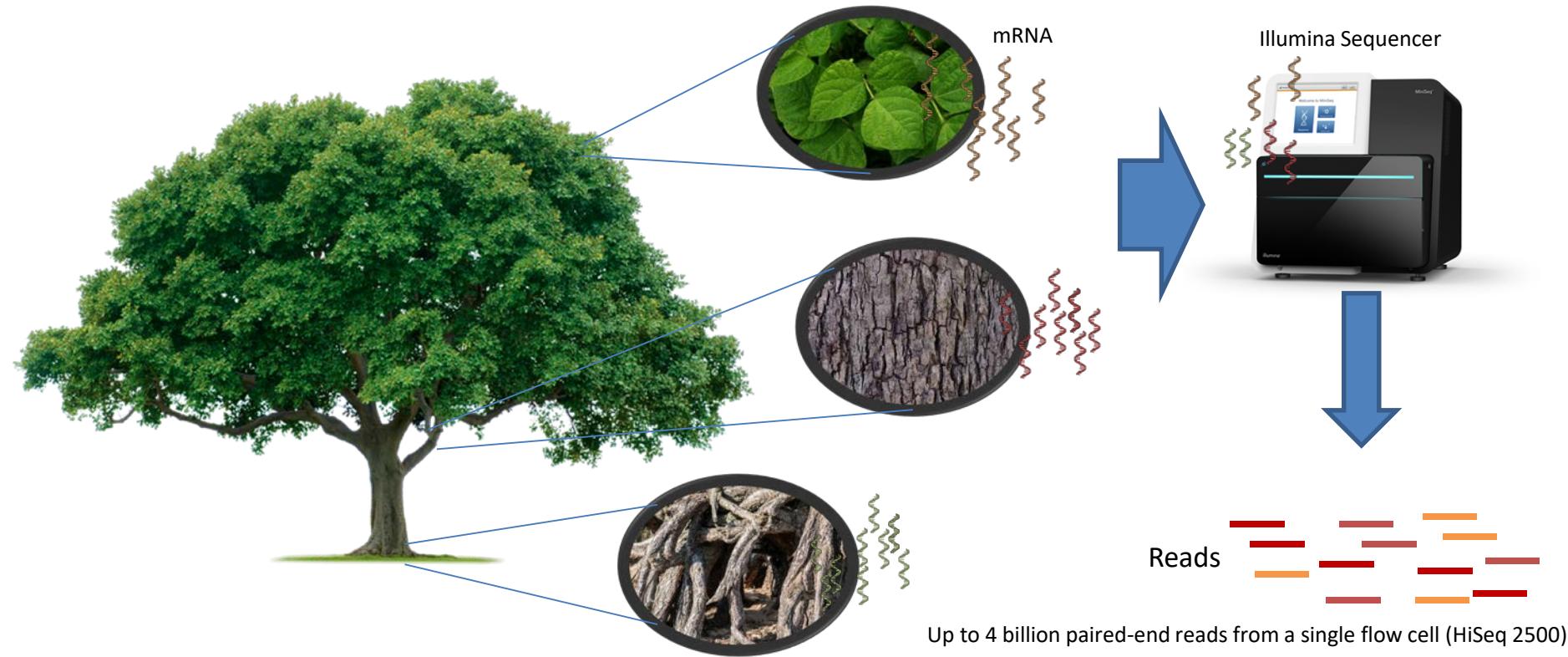
## Improving Genomes

Transcriptomes can be used to inform the gene space. The sugar pine genome was assembled using additional support from deep coverage RNA-seq data (Illumina and PacBio)

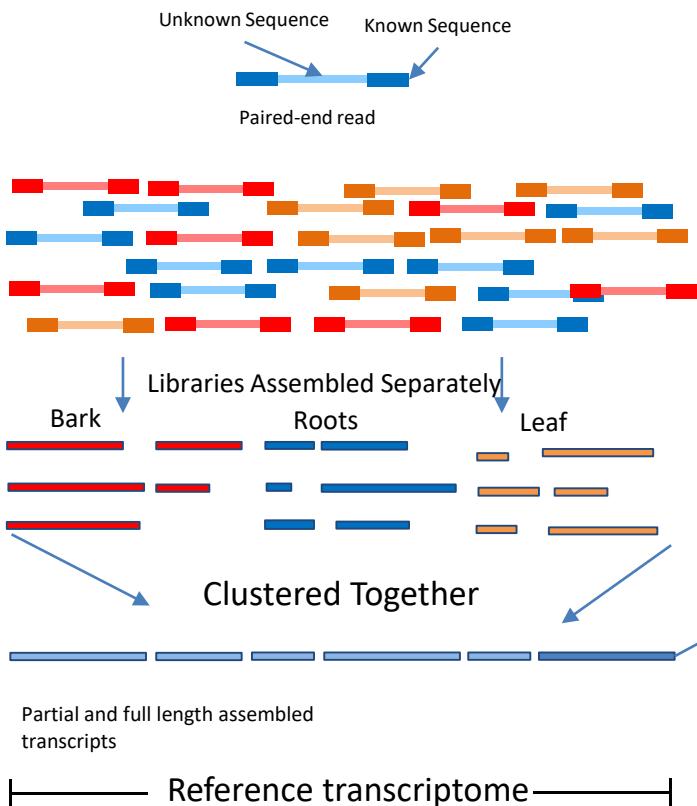
**UCD – Langley/Neale**



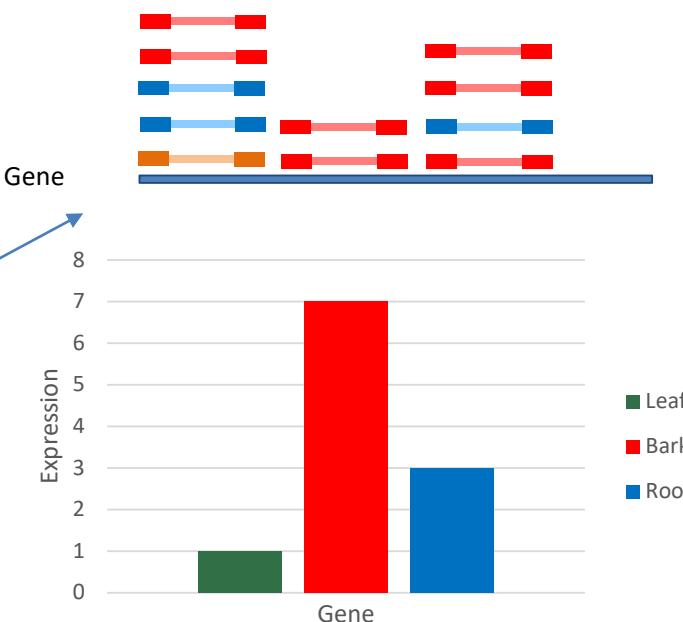
# Sample to sequence



## Transcriptome Assembly



## Mapping to Assembled Genes

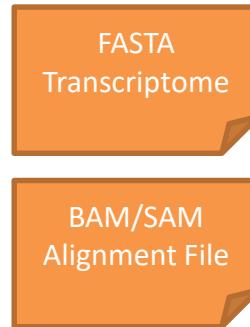


FASTA  
Transcriptome

BAM/SAM  
Alignment File

# EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline

- **Frame Selection** – GenemarkS-T
  - Provides information on complete, partial, and internal genes
- **Transcriptome filtering** - RSEM
  - Use BAM/SAM alignment file to filter transcripts based on expression values
- **Similarity Search** - DIAMOND
  - Best-hit selection based upon:  
*contaminants, scores, coverage, and phylogenetics, informativeness*
  - Leagues faster than traditional BLAST searching (*Butchfink et. Al 2015*)



- **Orthologous gene family assignment**– EggNOG
  - Assigns gene families
  - Applies relevant protein domains terms
- **Gene Ontology Annotation**
  - Incorporation of curated terms
  - Molecular function, biological process, cellular component
  - Leverages curated databases first
- **Output**
  - Statistics on hits, contaminants, databases, each stage in enTAP
  - Full annotated list in tab-delimited format



## Evaluating Frame Selection in Non-Model: Study Design

- Three non-model species: *Juglans regia* (Persian walnut), *Pseudotsuga menziesii* (Douglas-fir), and *Homalodisca vitripennis* (glassy-winged sharpshooter)
- Our study seeks to compare ORF detection methods across three different organisms with draft genomes. The organisms represent two plants (gymnosperm and angiosperm) as well as an insect.



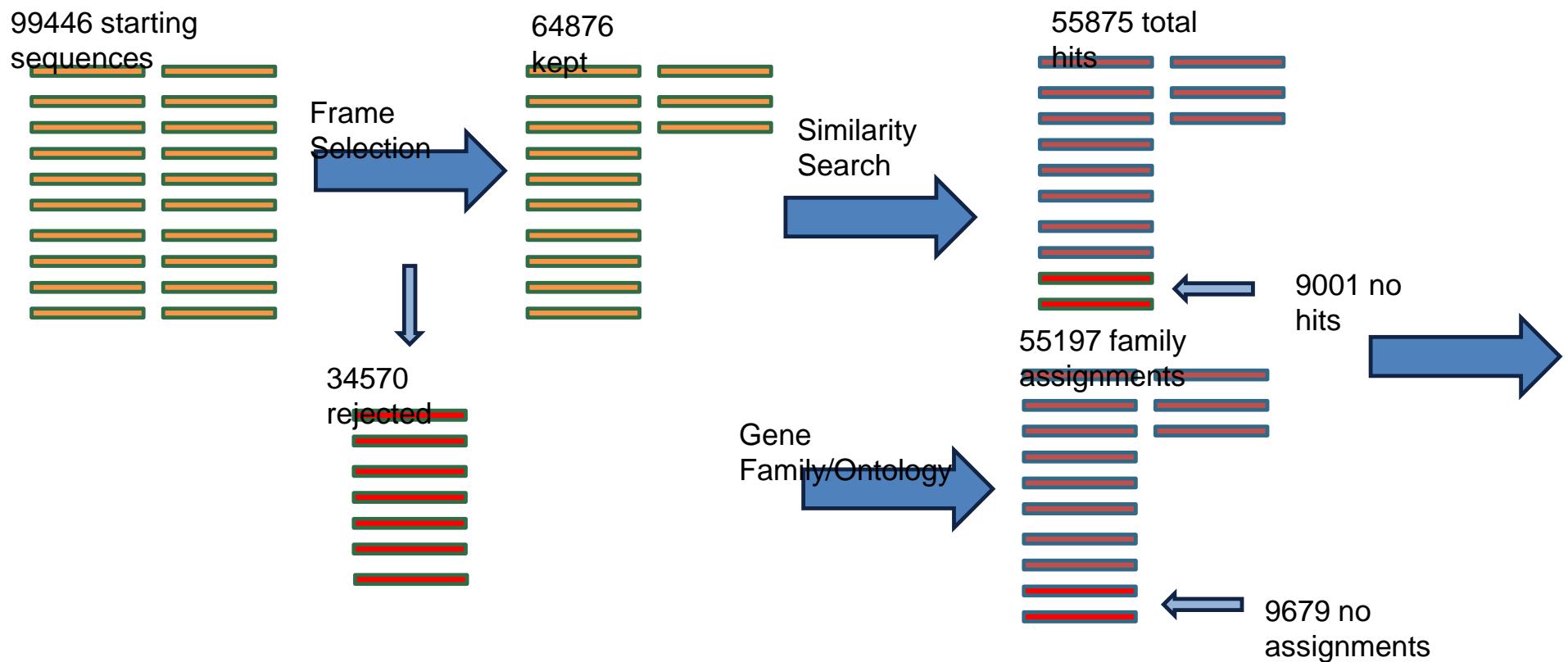
Species	Genome Size (Mbp)	N50 (bp)
<i>J. regia</i>	668	464,955
<i>H. vitripennis</i>	2200	776,706
<i>P. menziesii</i>	14500	387,073



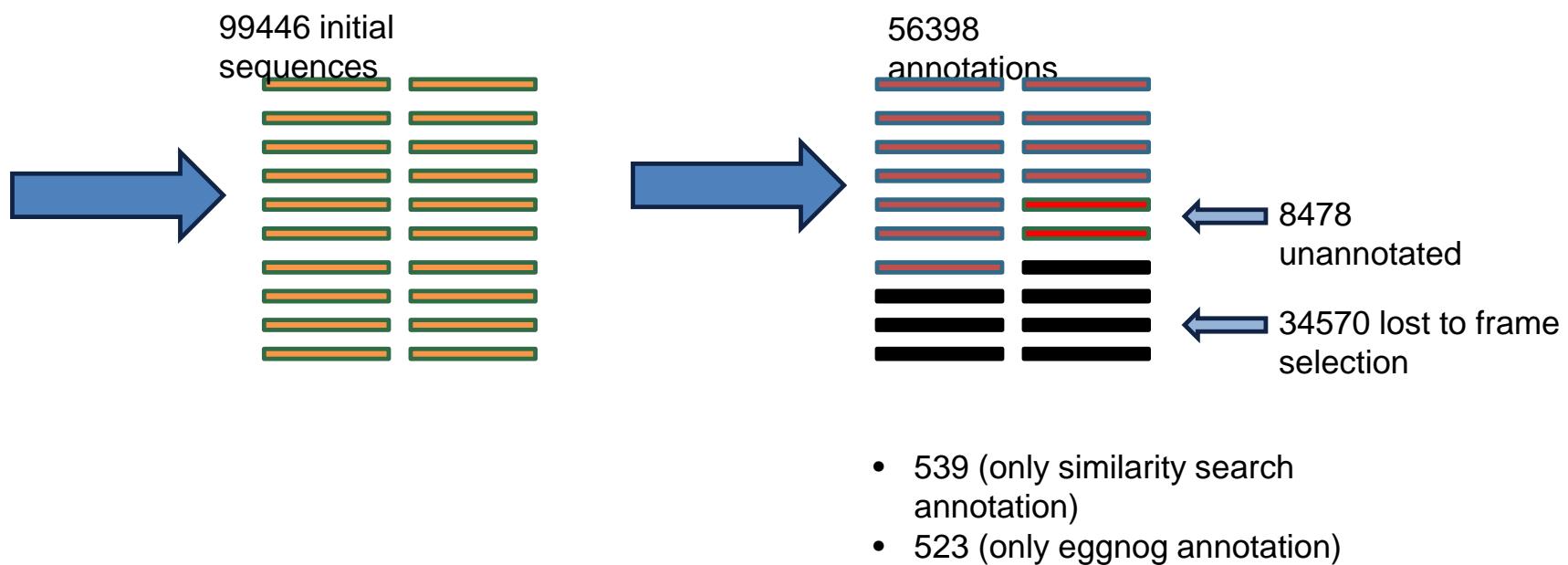
# EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline

- 100,000 sequences
  - Frame selection
  - Similarity Search
    - Uniprot Swiss-Prot
    - NCBI Refseq Complete
    - Arabidopsis
  - Eggnog
- Run time: 9hrs (8 cores)
  - Genemark: 80 min
  - Similarity Search: 406 min
    - Arabidopsis: 6 min
    - Refseq complete: 390 min
    - Swiss: 10 min
  - Eggnog: 150 min
  - enTAP ~30-45 min

# EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline

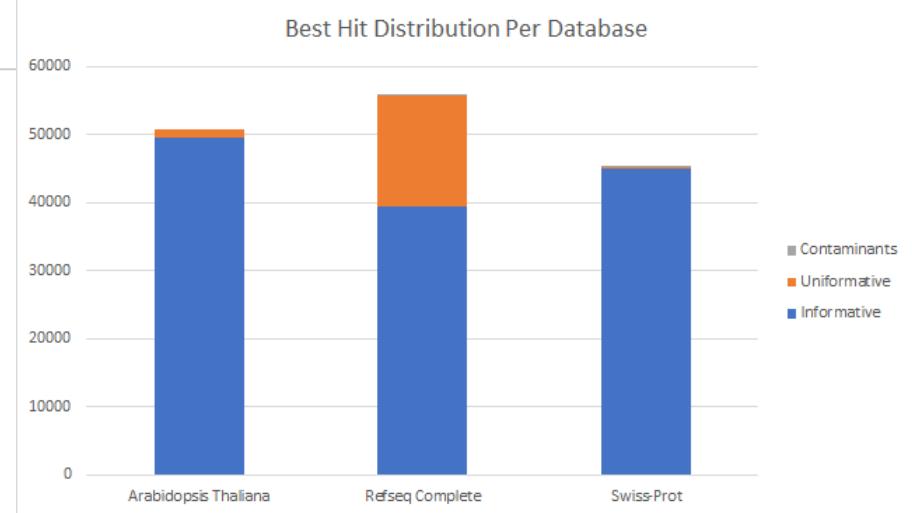
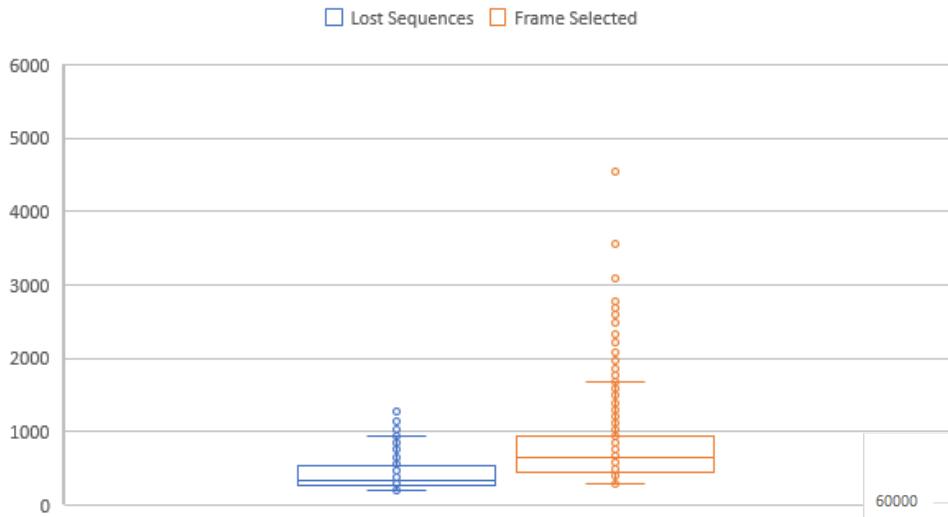


# EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline

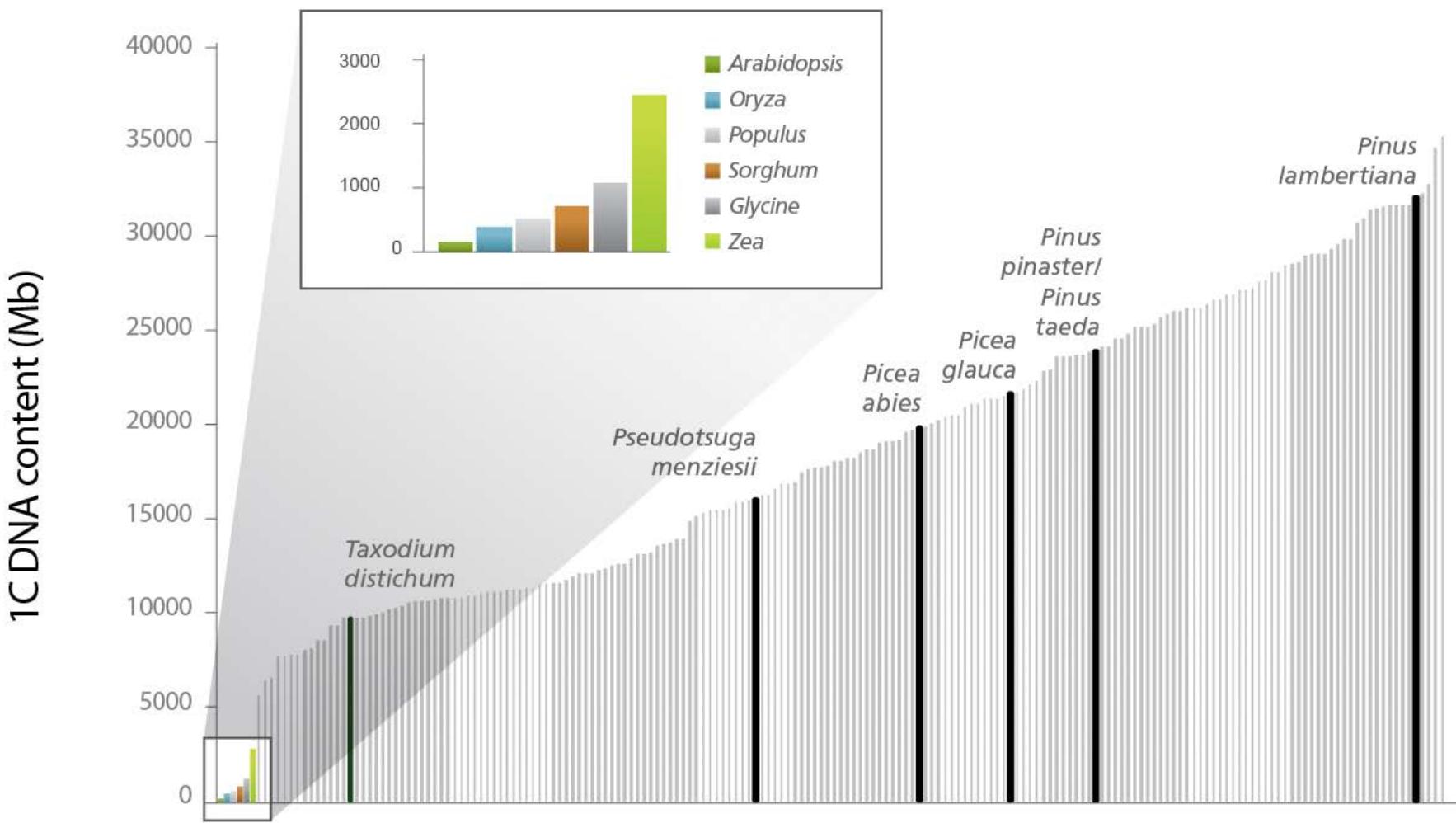


# EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline

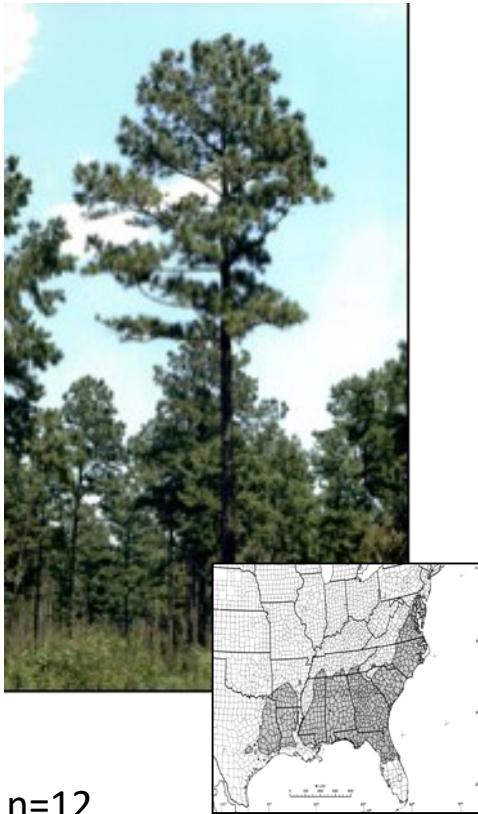
## GENEMARKS-T RESULTS



# PERSPECTIVE: CONIFER GENOMES



**Loblolly pine**  
(*Pinus taeda*)



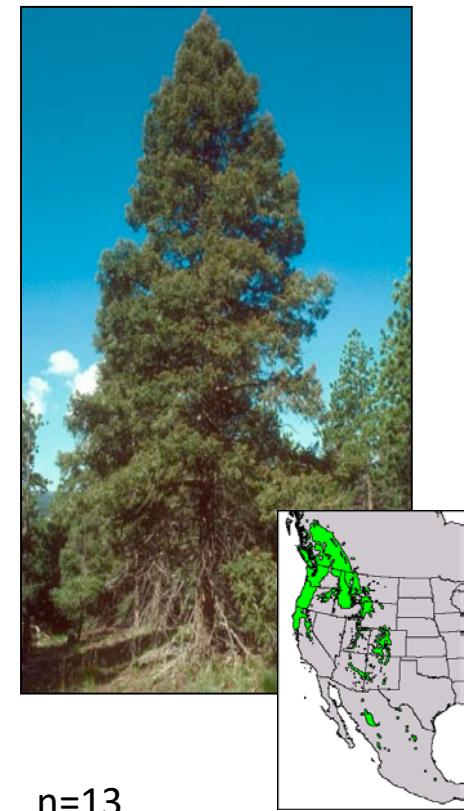
- n=12
- Genome size: 21.6 Gbp
- Genotype to sequence: 20-1010
- Mapping population: 6-1030x8-1070 and 20-1010x11-1060 (1000 F<sub>1</sub> progeny)

**Sugar pine**  
(*Pinus lambertiana*)



- n=12
- Genome size: 31.9 Gbp
- Genotype to sequence: 6000
- Mapping population: 5038x5500(1300 F<sub>1</sub> progeny)

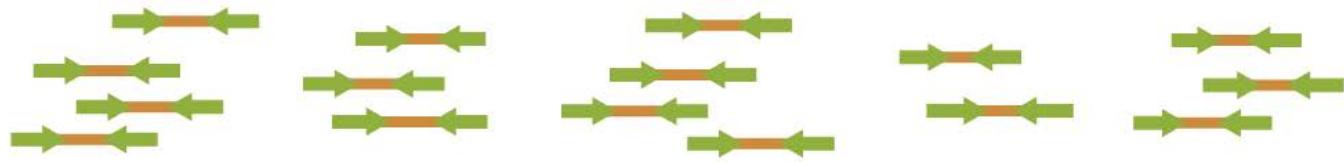
**Douglas fir**  
(*Pseudotsuga menziesii*)



- n=13
- Genome size: 18.6 Gbp
- Genotype to sequence: 412-2
- Mapping population: 412-2x013-1 (1000 F<sub>1</sub> progeny)

# ASSEMBLING THE REFERENCE GENOME (WGS)

Sheared genome  
fragments (200 to 600  
bp), prep and sequence  
using next-generation  
sequencing platform(s)

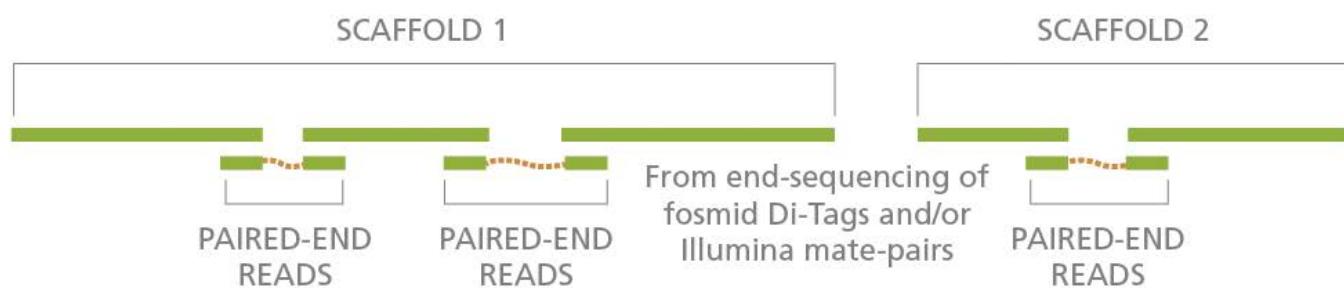


**16 billion paired reads ?!**

Continuous sequence  
– Contigs



Scaffold builds facilitated by paired-end or mate-pair reads

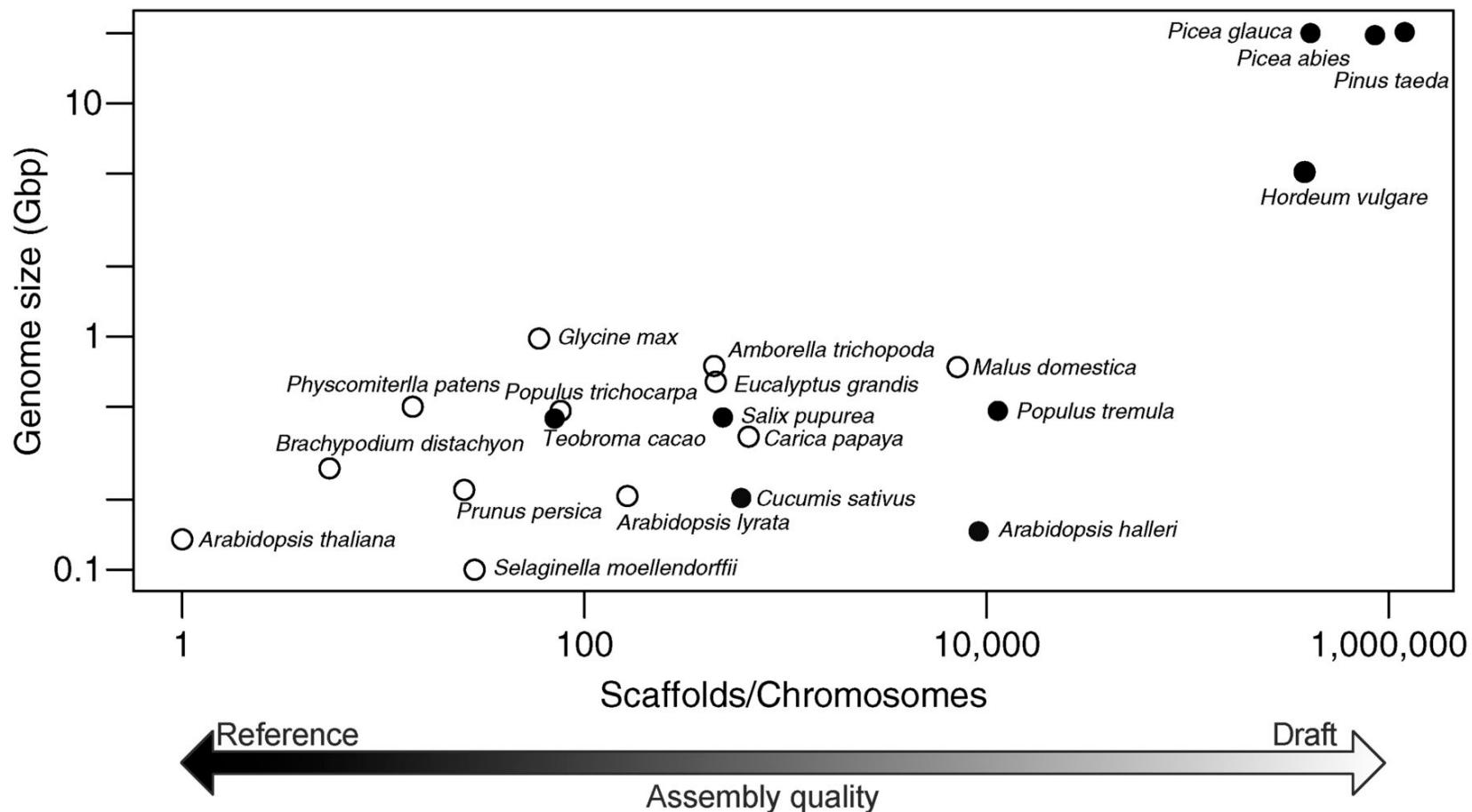


# Conifer Genomes Compared

\*Includes transcriptome scaffolding for all three genomes with existing/new resources

Species	<i>Pinus taeda</i> (v1.01)	Ptaeda 2.0
Estimated genome size (Gbp)	<b>21.6</b>	<b>21.6</b>
Total scaffold span	<b>22.6</b>	<b>22.1</b>
N50 contig size (Kbp)	<b>8.2</b>	<b>25.3</b>
N50 scaffold size	<b>66.9</b>	<b>107.1</b>
Number of scaffolds	<b>9,412,985*</b>	<b>1,496,869</b>
Assembler	<b>Masurca</b>	<b>Masurca</b>

# Genome Assemblies Compared



# Genome annotation

**~32 billion bp**

ACAATAAATCACATTAATTCTTATCTCATGTGAAATTTCATATTATGATT  
GATACCTTAAATGTCATTGTTGAAGGAAGATTATTCACTTTTCATTCA  
ATAAATATTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTT  
CTAGGCATTGGGATACCATGTTACAAGACAGACTATGATTACAGGA  
TCAGATGTGGACTCTCAAATTGACTGAGAATAAAACAGACACTAAACAA  
GTAATAAAGTTAATTCAAGTTGAATTGATGCTAGAAAGACAATGAAAC  
AGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTT  
AGATAAGGTACCTGATTGGTGGATTGGAAGACCTCTGAGAGATTAGTG  
TCTTCAGATATGCCATTGATGAAAGAACATTCACTGGGAAGGCCTA  
GCATTAaaaACCGCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCTG  
GATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAA  
AATGAAAATAGATTTAAACATGTTAATTACGTTACTTTTGTAAATT  
ACTTTCTCTTCACTCTTACCTGTCAATGTTATTAAATATTTTAGGAAC  
AATAAATCACATTAATTCTTATCTCATGTGAAATTTCATATTATGATTGA  
TACCTTAAATGTCATTGTTGAAGGAAGATTATTCACTTTTCATTCAATA  
AATATTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTA  
GGCATTGGGATACCATGTTACAAGACAGACTATGATTACAGGATCA  
GATGTGGACTCTCAAATTGACTGAGAATAAAACAGACACTAAACAAGTA  
AATAAAGTTAATTCAAGTTGAATTGATGCTACTATGGAAAAATGAAAAT  
AGATTTAAAACATGTTAATTACGTTACTTTTGTAAATTACTTTCTT  
CTTTCACTCTTACCTGTCAATGTTATTAAATATTTTAGGAACAATAAATCA  
CATTAATTCTTATCTCATGTGAAATTTCATATTATGATTGATACCTTAA  
ATGTCATTGTTGAAGGAAGATTATTCACTTTTCATTCAATAAAATATTTT  
TAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGG  
GGATACCATGTTACAAGACAGACTATGATTACAGGATCAGATGTGGA  
CTCTCAAATTGACTGAGAATAAAACAGACACAAACAAGTAAATAAGTT  
AATTCAAGTTGAATTGATGCTATCCCAGGCACAAGACCA....

## Genes:

- Coding, noncoding, miRNA, etc.
- Isoforms
- Expression

## Genetic variation:

- SNPs

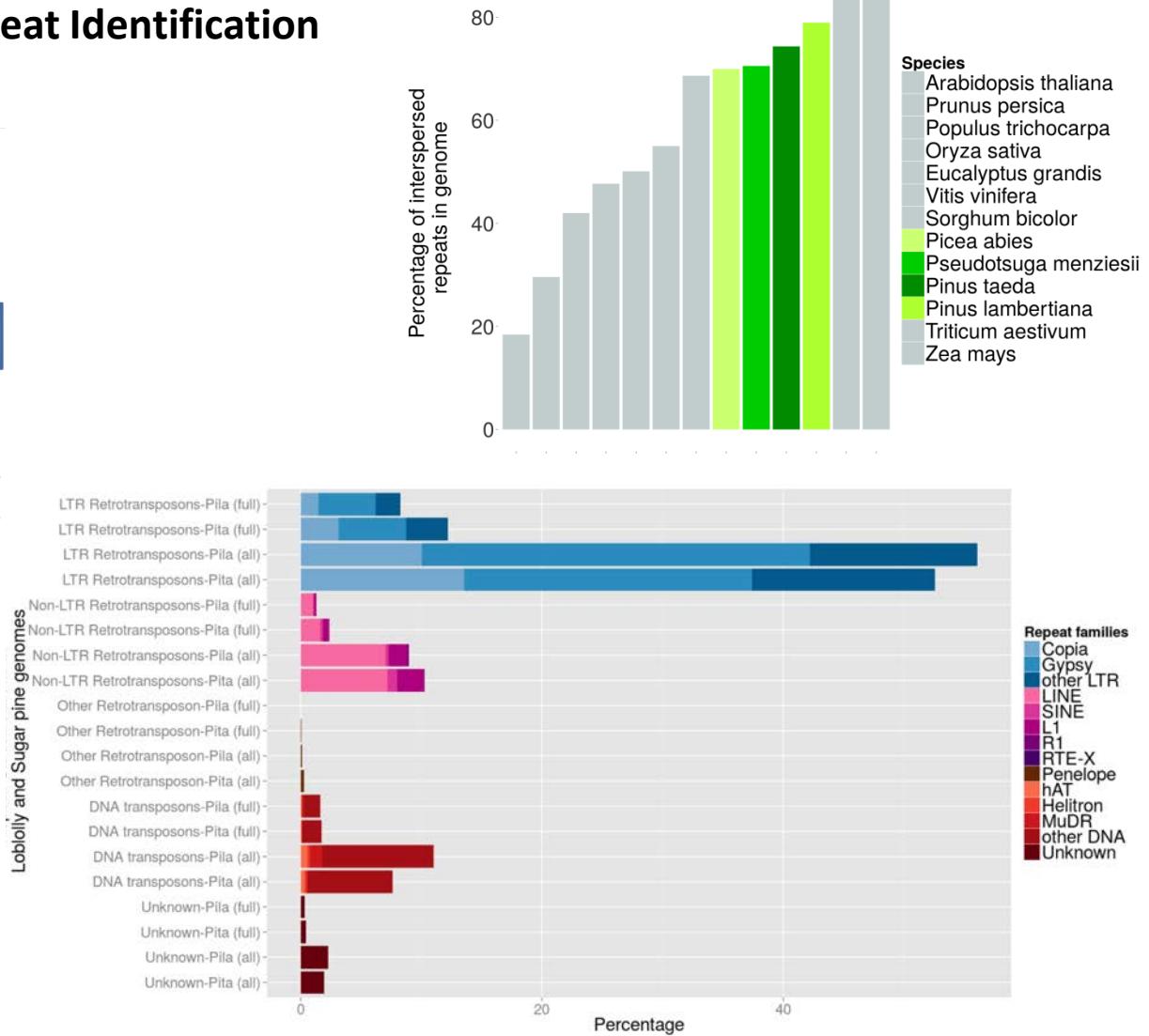
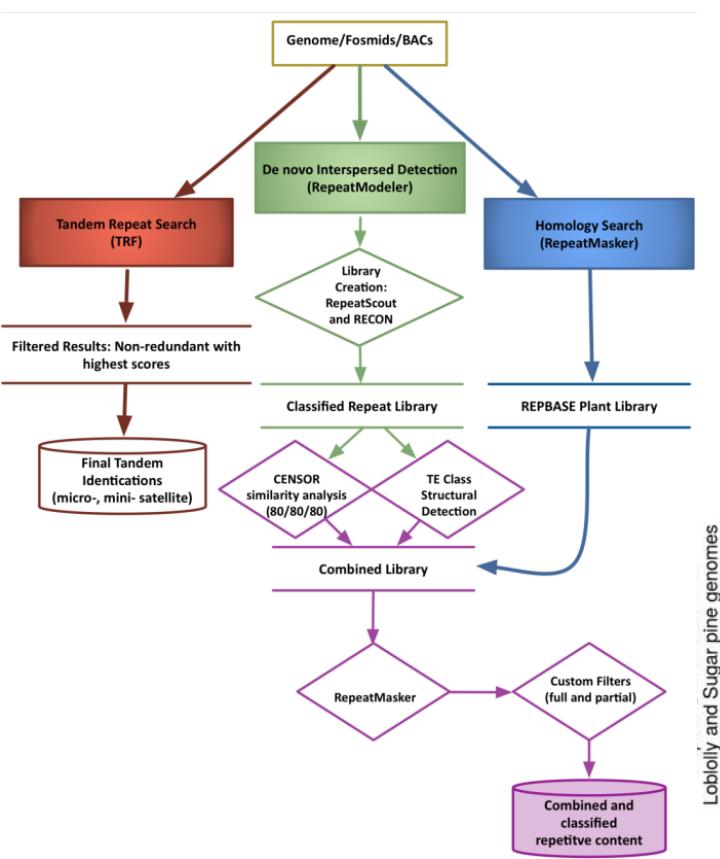
## Regulatory sequences:

- Promoters
- Enhancers

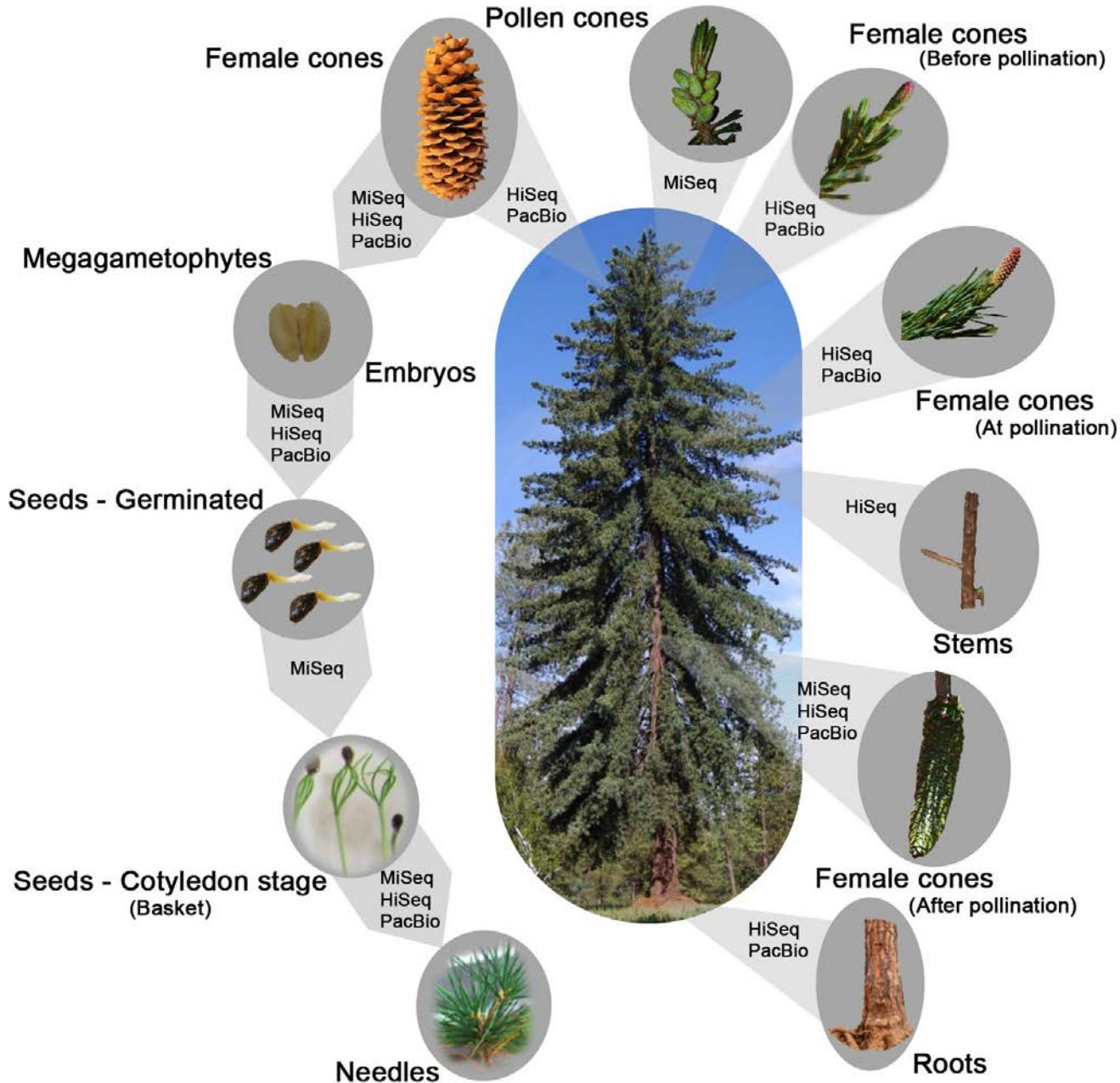
## Epigenetics:

- DNA methylation
- Chromatin

# Similarity and *de novo* Repeat Identification



	loblolly pine fosmids	sugar pine fosmids	Douglas-fir fosmids	sugar pine WGS	loblolly pine WGS
Length of genome (Mbp)	277	160	117	$24.7 \times 10^3$	$17.8 \times 10^3$
% of interspersed repeat content	80.2	76.6	72.7	88.96	84.37



# Genome Annotation: Genes!

Cyverse (TACC)

Ran in 72 hours on 8,000 cores

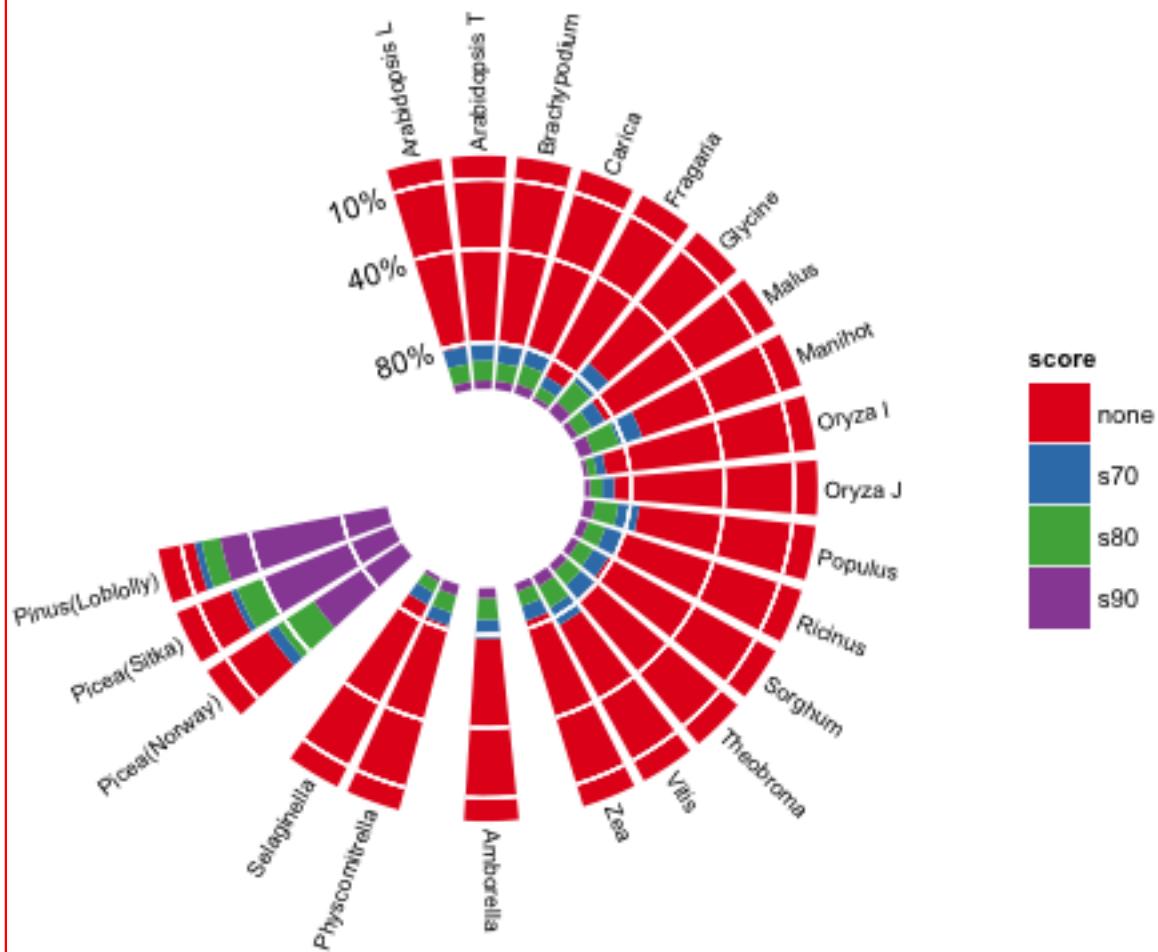
Provided *ab initio* gene predictions  
for an additional 22,345 full length  
genes.

29,189 de novo transcriptome +  
42,345 unique additions =

**> 71,534 genes (round 1)**  
**> 8,000 genes (round 2)**

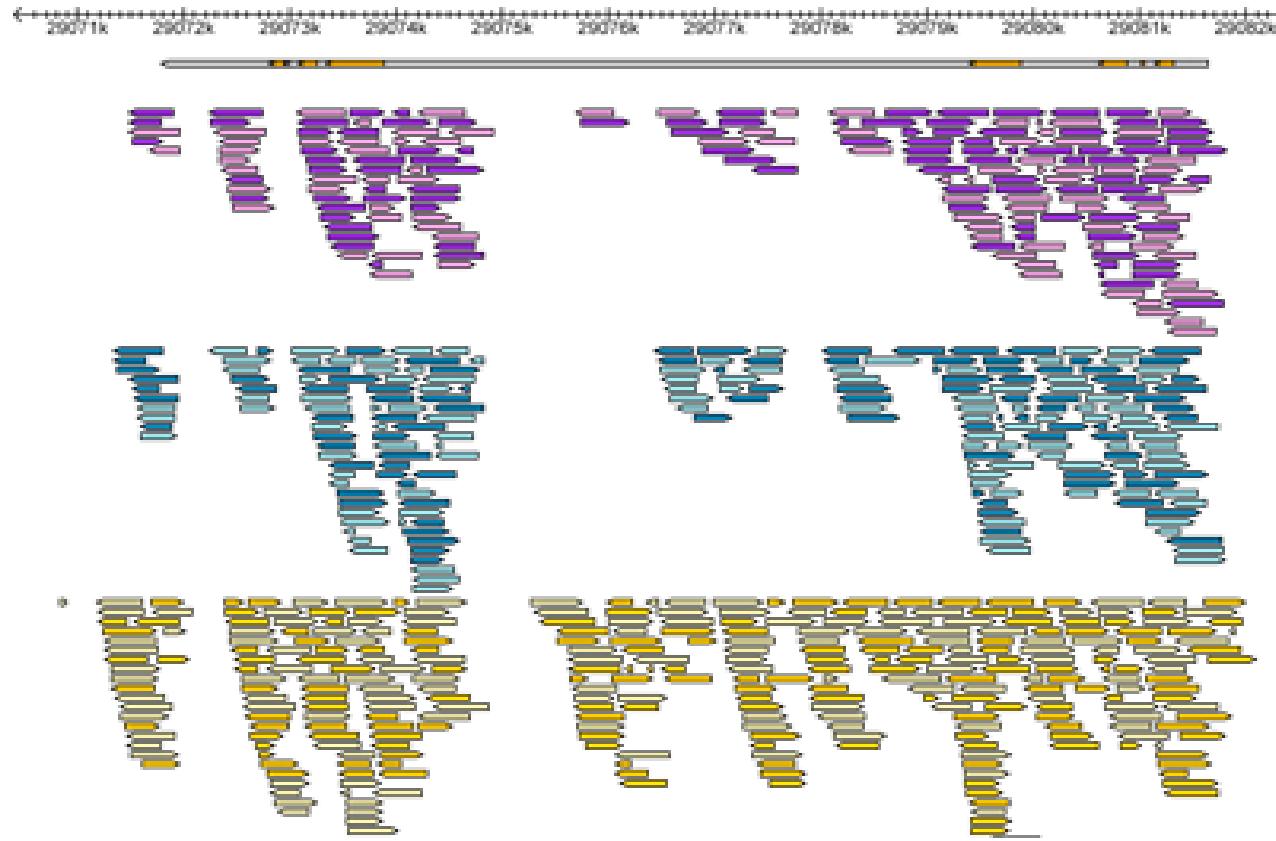
## Challenges:

- Fragmented genome
- Pseudogenes
- Transcriptomic assemblies
- -> Overall poor gene models



# Improving Gene Annotation

## Model developed on walnut (*Juglans*) genomes

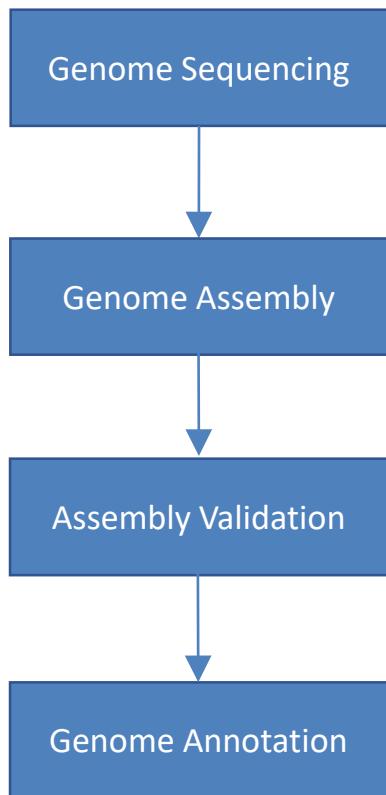


Masking + RNA-Seq Reads + Pseudogene + Assembled Evidence

**loblolly pine 2.01 - 33,215 gene models**

Run time: 2 days on 64 cores

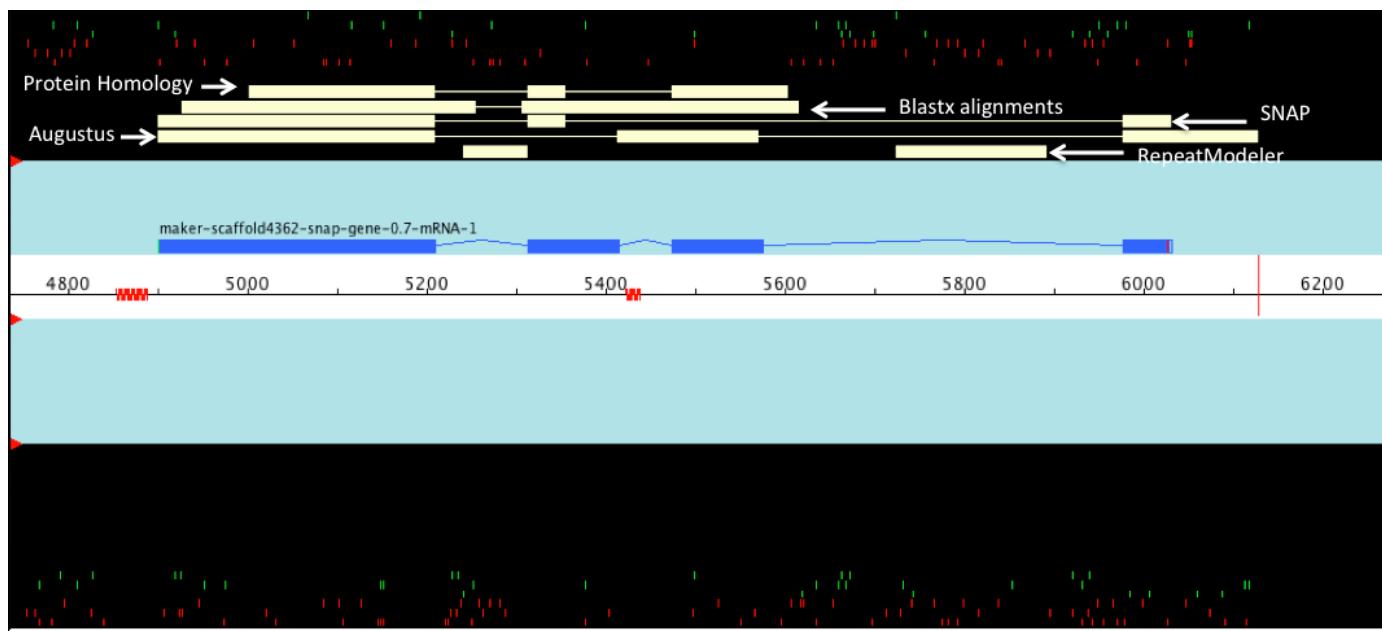
# Annotating *Juglans regia* (Common Walnut)



- Genome sequenced using HiSeq 2500 (Illumina)
- Two different assembly methods
- Transcriptome scaffolding
- Creation of Pacbio data
- Tandem & interspersed repeat identification
- Gene space completeness through **MAKER**

# MAKER: An Annotation Pipeline

- The Maker pipeline leverages existing software tools and integrates their output to produce the best possible gene model for a given location based on alignment evidence.

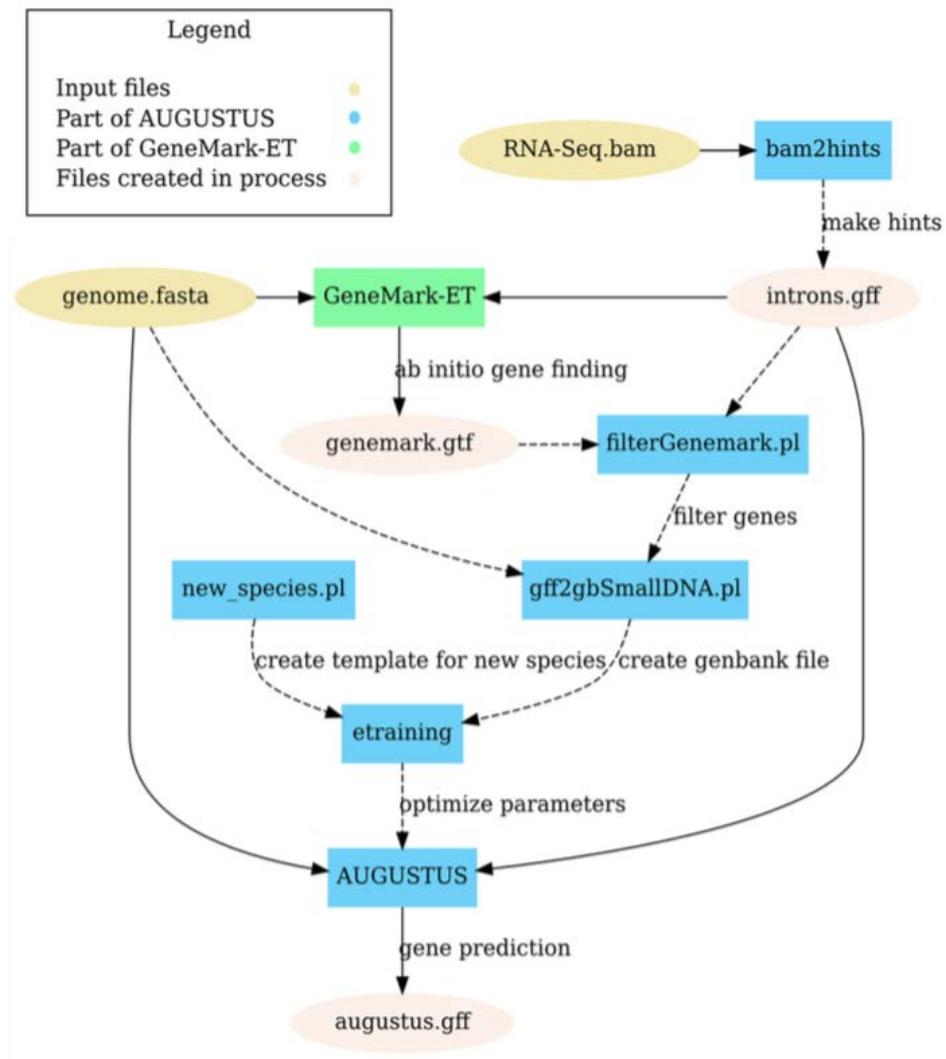


# MAKER Annotation Results

- Overall number of gene models = 32,496
  - Classify these genes as high quality completes, high quality partials, and low quality.
  - ***High Quality Completes ~ 52%***  
High Quality Partials ~ 27%  
Low Quality ~ 21%
- Limitations of MAKER
  - Uses NCBI BLAST to calculate alignment evidence
  - Requires training gene predictors like Augustus
  - Requires compiling a lot of evidence as input for accurate gene models
- Alternative?

# BRAKER: Another Annotation Pipeline

- Solely relies on two software:
  1. Augustus
  2. GeneMark-E-T
- Requires only two inputs:
  1. Assembled Genome
  2. Alignments of raw RNA reads to assembled genome
- Pipeline developed in-house to combine aspects of BRAKER with EvidenceModeler



# BRAKER/EvidenceModeler Annotation Results

- Overall number of gene models = 146,465
  - High quality set of genes:
    1. Complete canonical multiexonic genes with a valid protein domain = 42,772 genes
    2. Complete monoexonics genes aligning to “monoexonic gene database” = 343 genes
- Validation of High Quality Multiexonic Genes
  - EnTAP annotation
    - 41,472 genes aligned to Refseq Plant Protein or Uniprot Database with coverage > 50%.
    - Leaves 1,300 genes unaccounted for → confirmed to be ‘walnut specific’
- Further Validation
  - Captures ~75% of MAKER genes
  - Validated against transcriptome

# Acknowledgements



## USDA Agricultural Research Service

Brian Knaus

## Utah State University

Hardeep Rai

## Washington State University

Doreen Main

Stephen Ficklin

## University of Tennessee

Meg Staton

## Clemson University

Alex Feltus

## North Carolina State University

Fikret Isik

John Frampton

Ross Whetten



**JOHNS HOPKINS**  
MEDICINE



## University of Maryland

Aleksey Zimin

James A Yorke

## Texas A&M University

Carol Loopstra

Jeffrey Puryear

Claudio Casola

## Johns Hopkins University, School of Medicine

Daniela Puiu

Steven L. Salzberg

## Indiana University

Keithanne Mockaitis

## USDA Forest Service

Detlev Vogler

Camille Jensen

Annette Delfino-Mix

Jessica Wright

Richard Cronn



## University of Connecticut

Ethan Baker

Taylor Falk

Uzay Sezen

Gaurav Sablok

Nic Herndon

Daniel Gonzalez-Ibeas

Robin Paul

Steven Demurjian, Jr.

Emily Grau

Alex Hart

Qiaoshan Lin



## University of California, Davis

David Neale

John Liechty

Pedro J. Martinez-Garcia

Patricia Maloney

Randi Famula

Hans Vasquez-Gross

Charles H. Langley

Kristian Stevens

Marc Crepeau

## University of Colorado



University of Colorado  
Boulder | Colorado Springs | Denver | Anschutz Medical Campus



## University of Colorado, Boulder

Jeffry Mitton

## University of California, Merced

Lara Kueppers