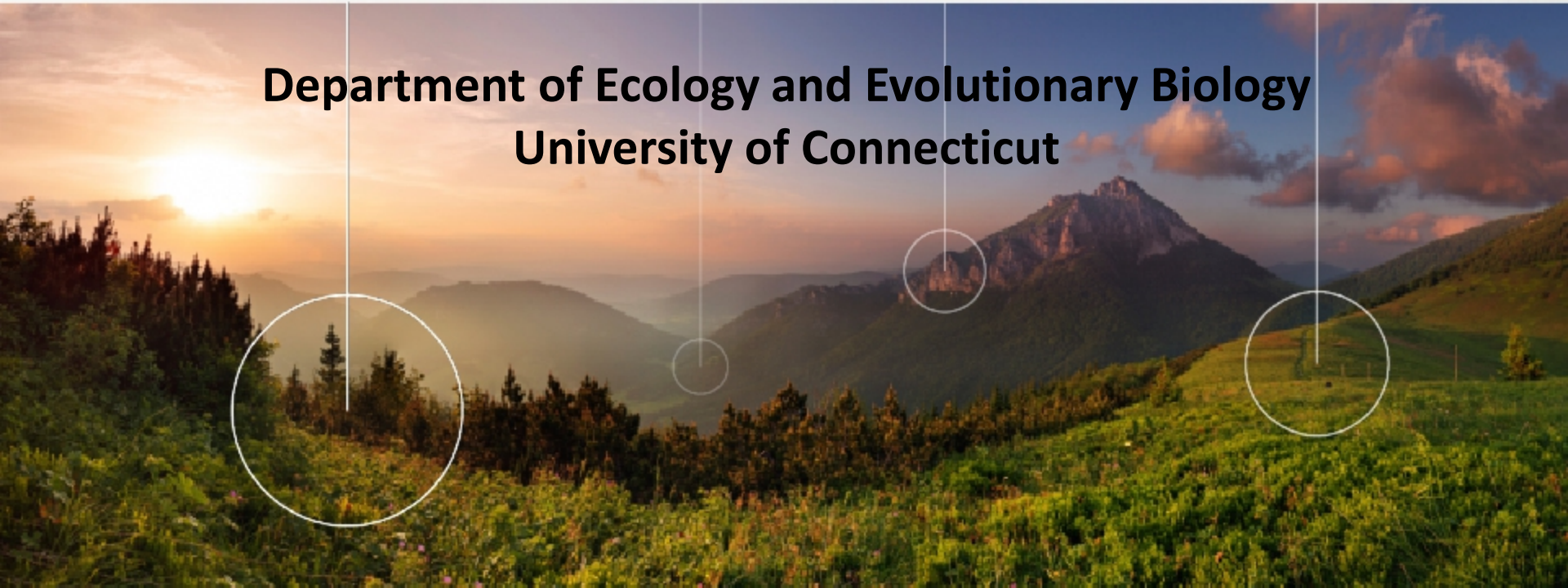
Computational approaches to decode megagenomes and develop database resources for the forest tree community

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."

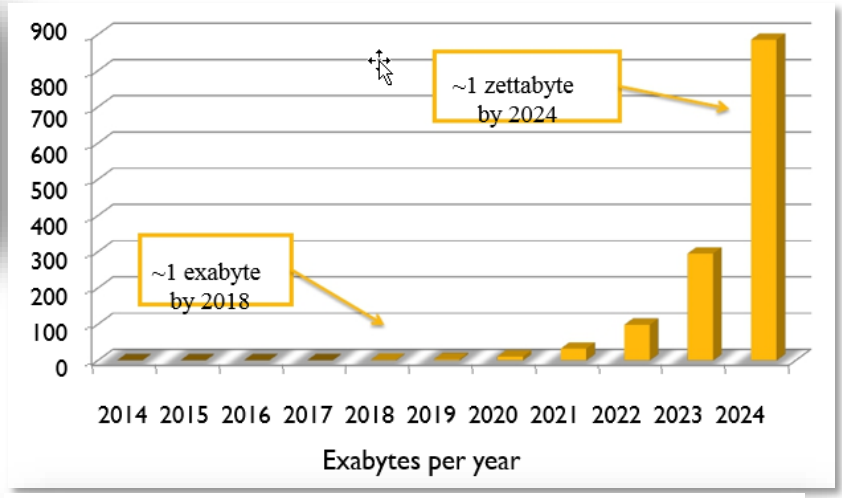
Big Data in Genomics

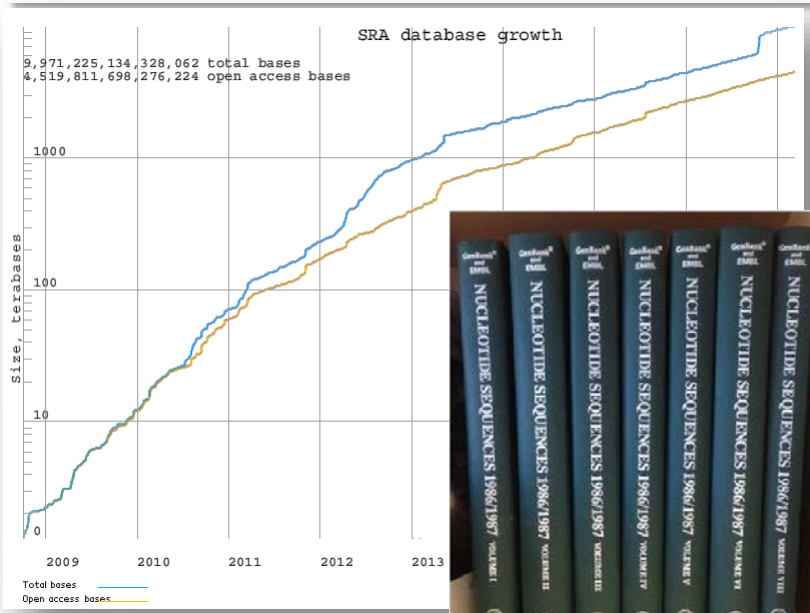
PERSPECTIVE

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



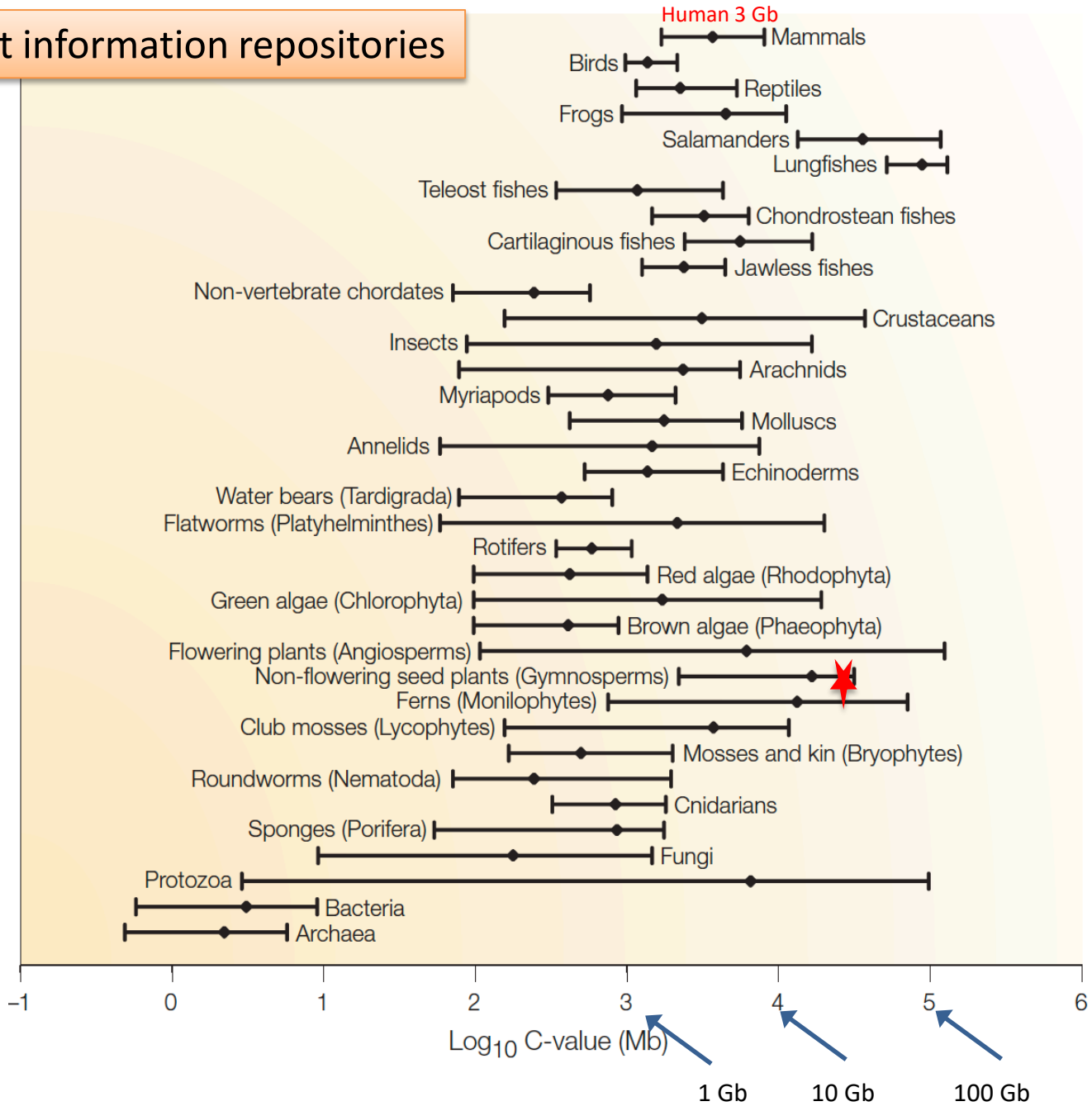


"Compared genomics with three other major generators of Big Data: ***Astronomy, YouTube, and Twitter...***Genomics is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis"

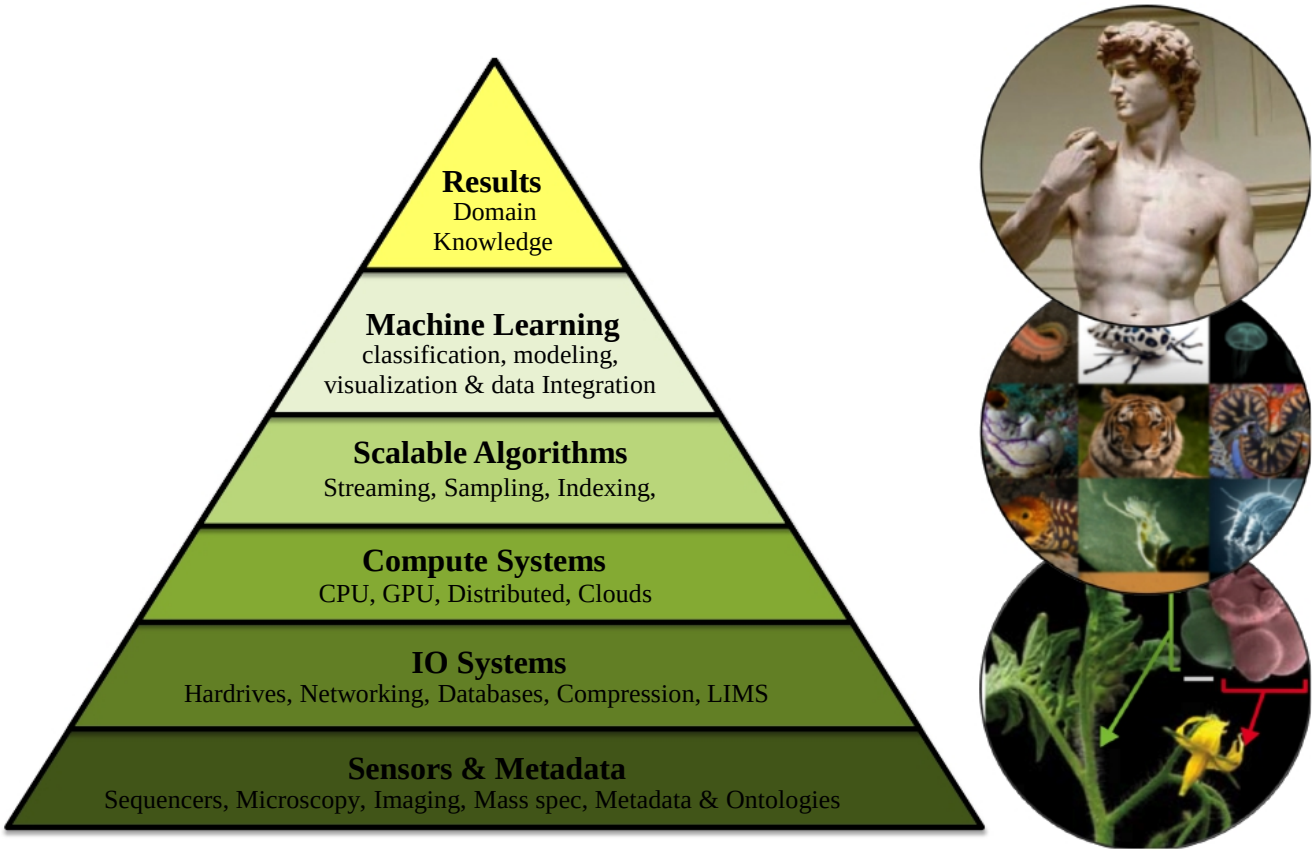
Mostly Genomic but…Proteomics, Phenomics, Metabolomics…

Genomes are vast information repositories

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



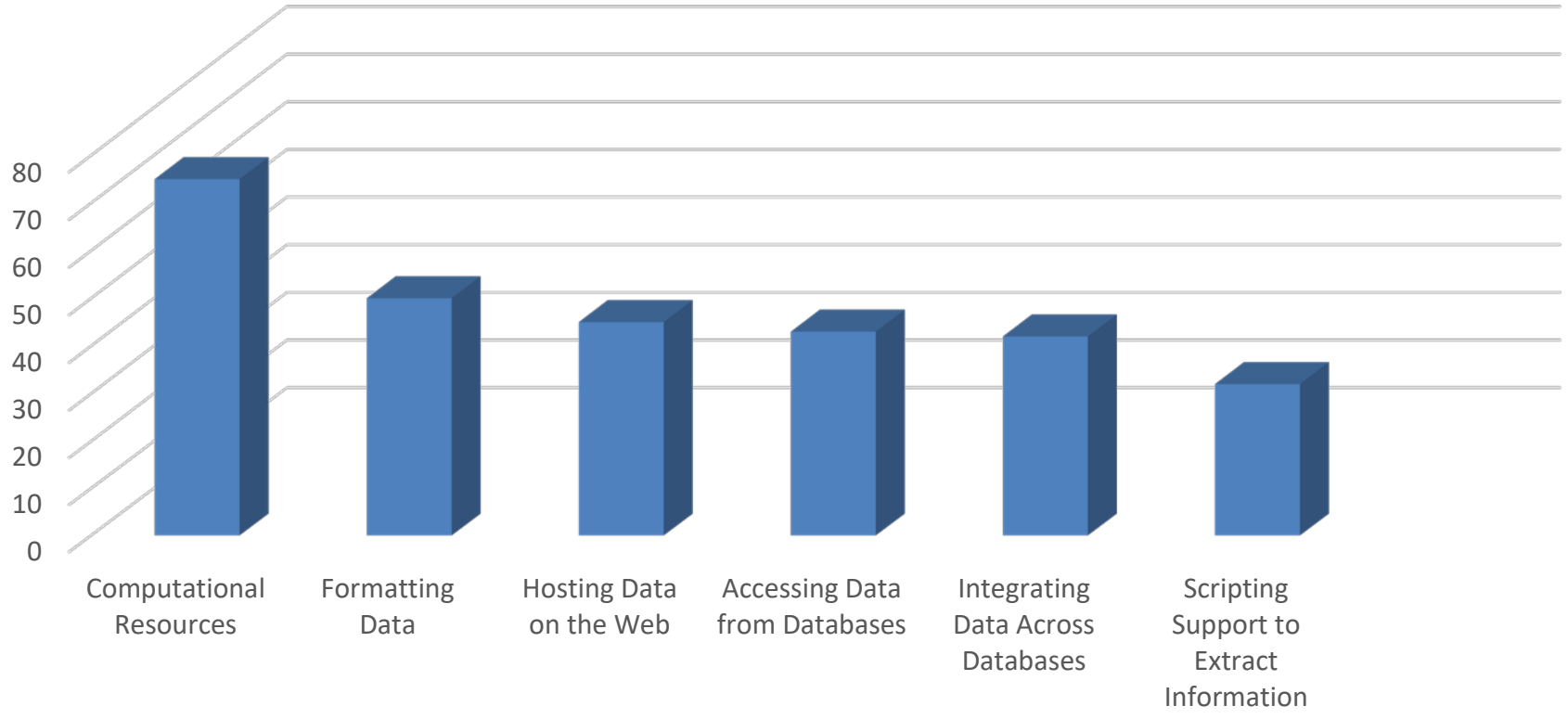
Acquiring Knowledge through Big Data



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 (~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

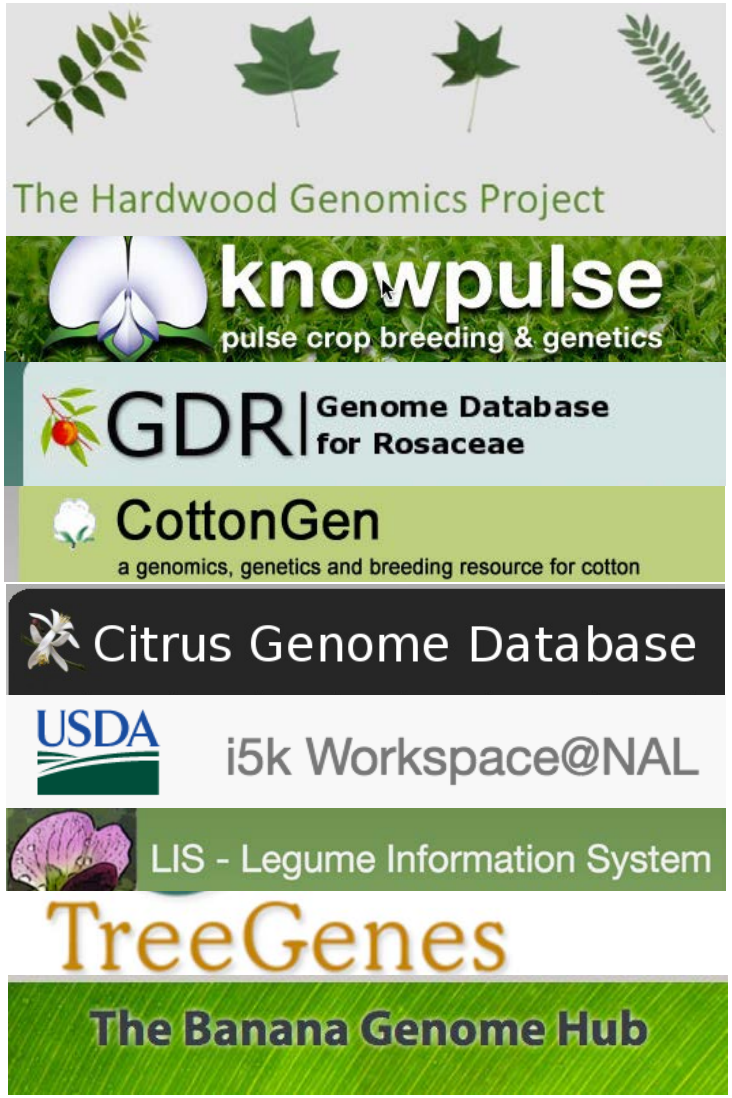
Tripal

Open source content management system (CMS) for biological data

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

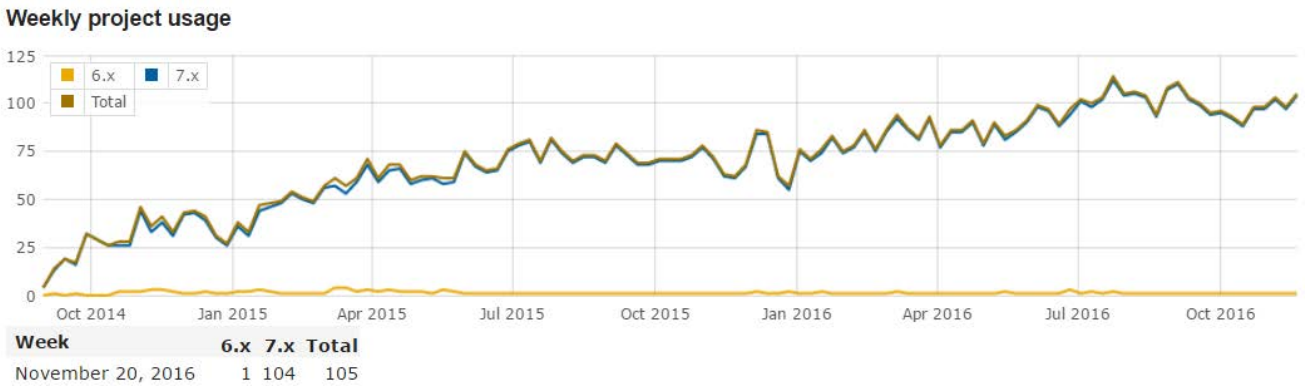
Benefits:

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



Current State of Tripal

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



TREEGENES DATABASE

TreeGenes
A Forest Tree Genome Database

Welcome | Research | TreeGenes | DiversiTree | FTGSC | iPlant | Resources | Events | News | Jobs | Links | Help

Welcome to the TreeGenes Project!

TreeGenes

Sequence Resources
Summary by Genus

Colleague Directory
Colleagues
Organizations

Species Database
Forest Trees

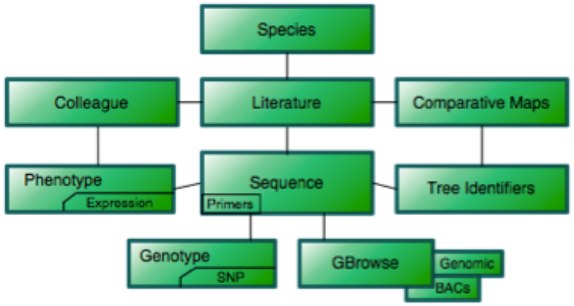
Literature Database
Search Literature

Transcriptome Database
Search Transcriptome
Transcriptome Summary

Protein Database
Search Proteins
Protein Summary

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

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

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

NOME

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

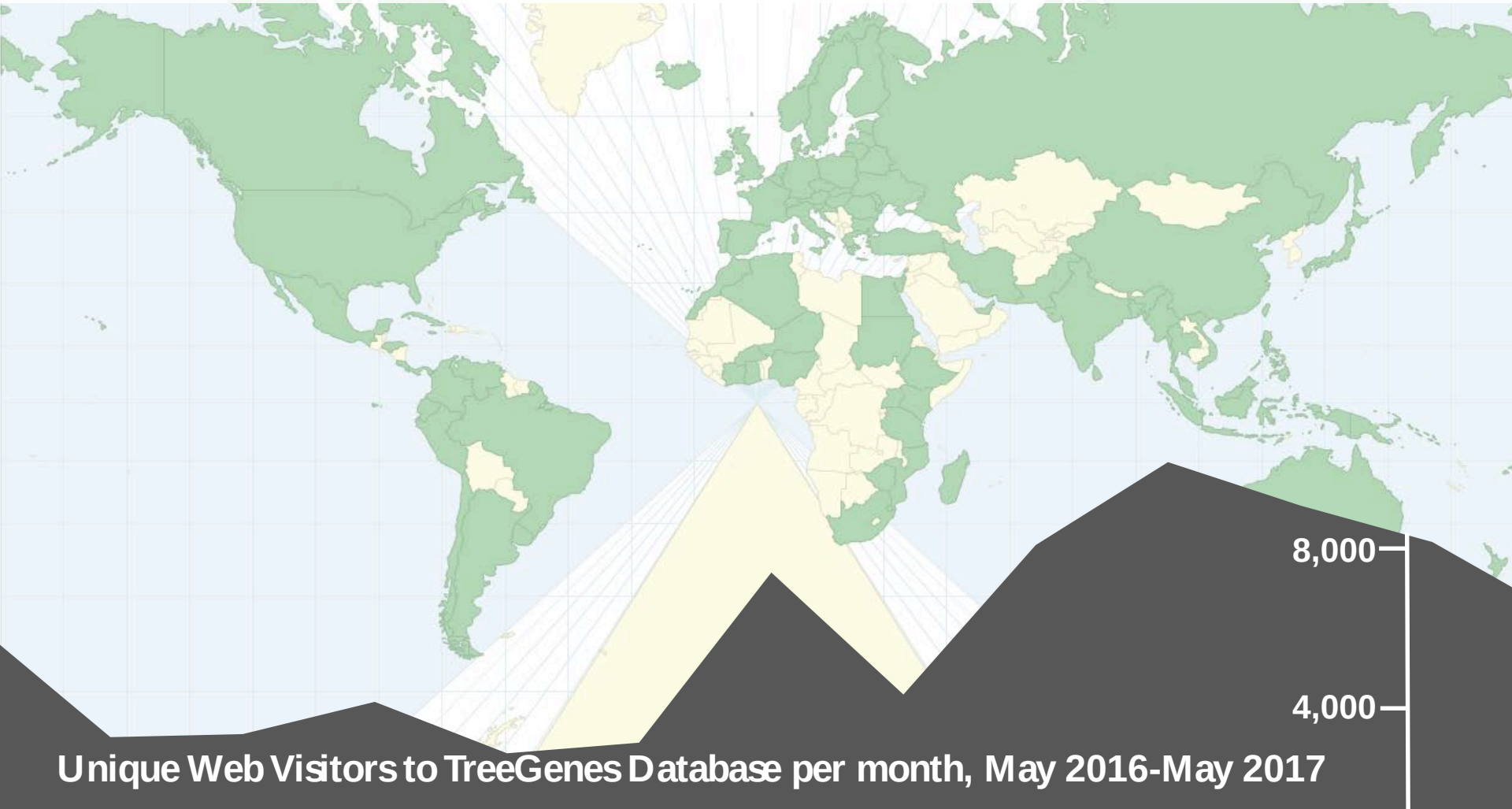
treegenesdb.org

- 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)

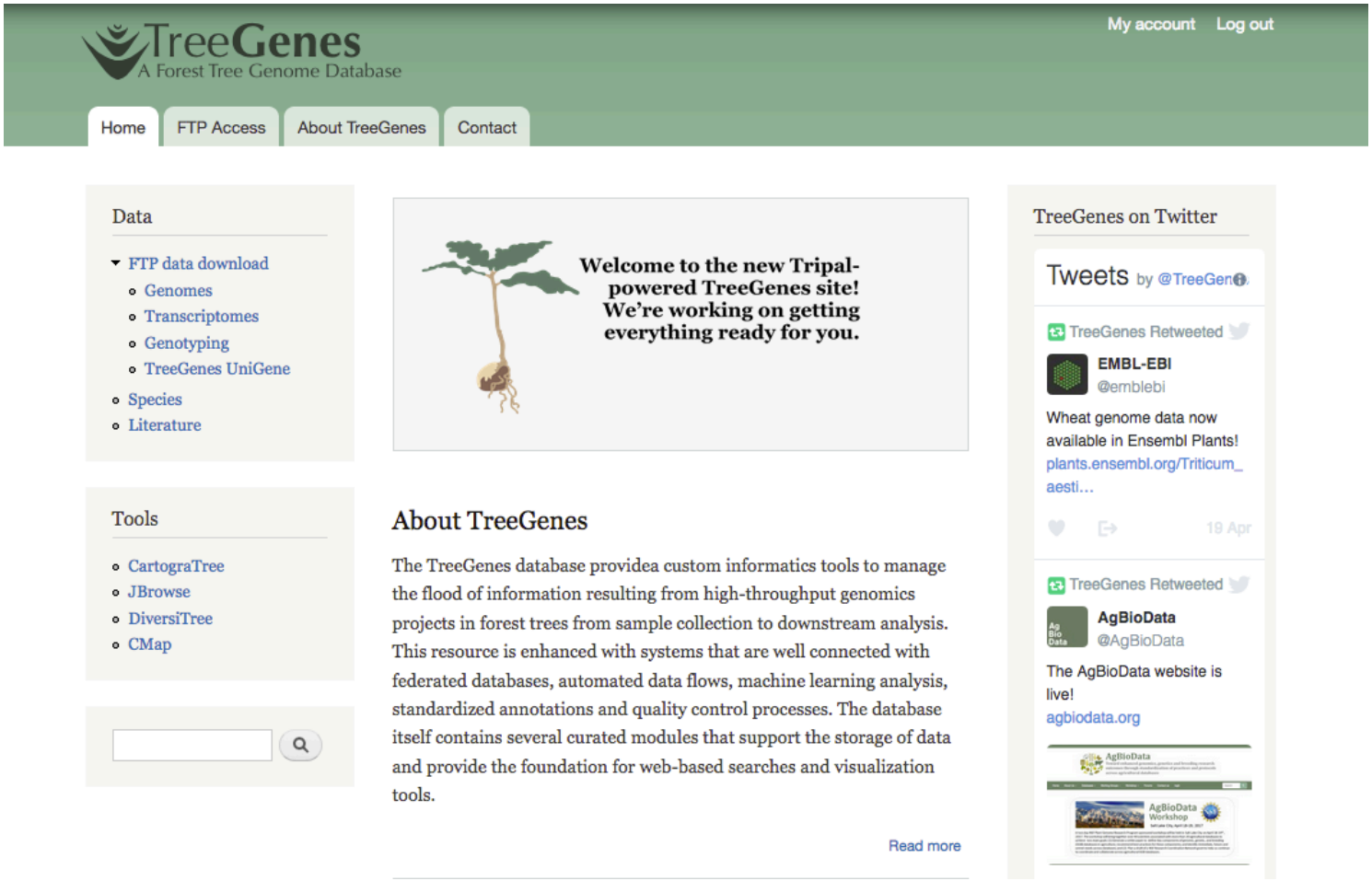
TreeGenes Database: Users

treegenesdb.org

2,012 users from 855 organizations in 92 countries

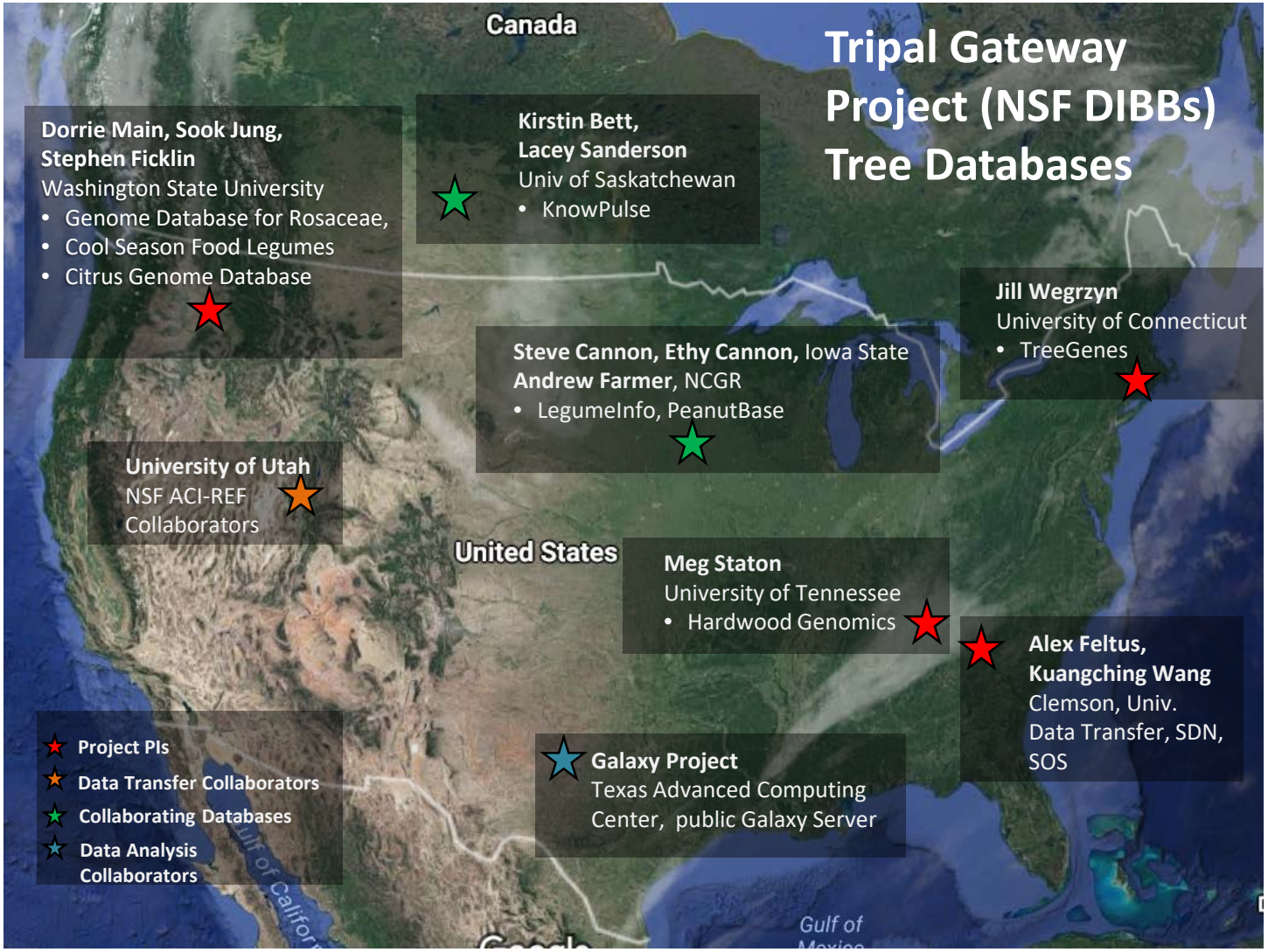


New TreeGenes Coming Soon!

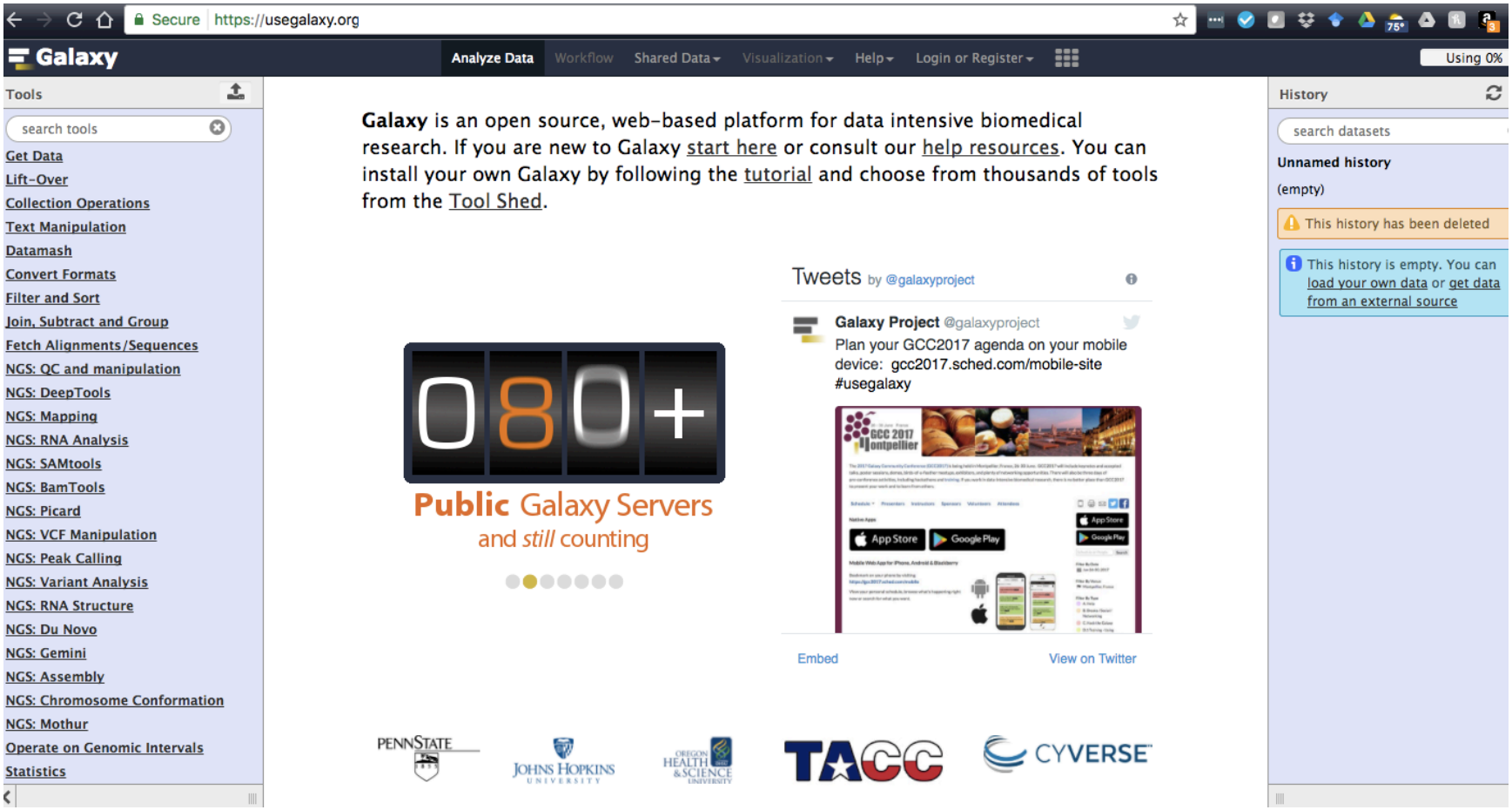


Tripal Gateway Project (Data Provider)

- 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 (NSF DIBBs) Tree 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

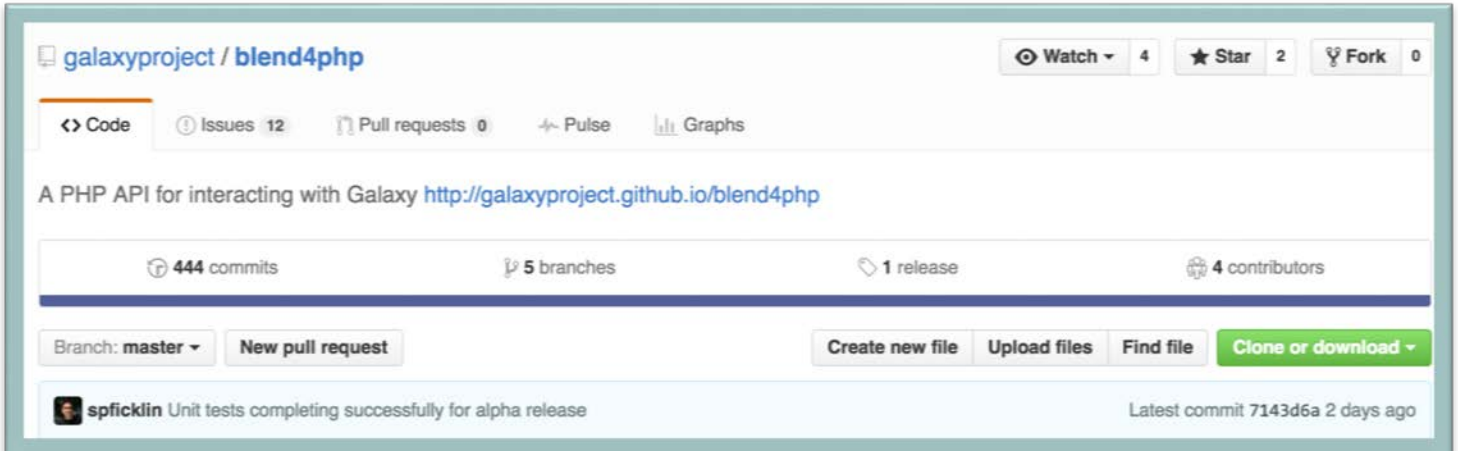
What is Galaxy?



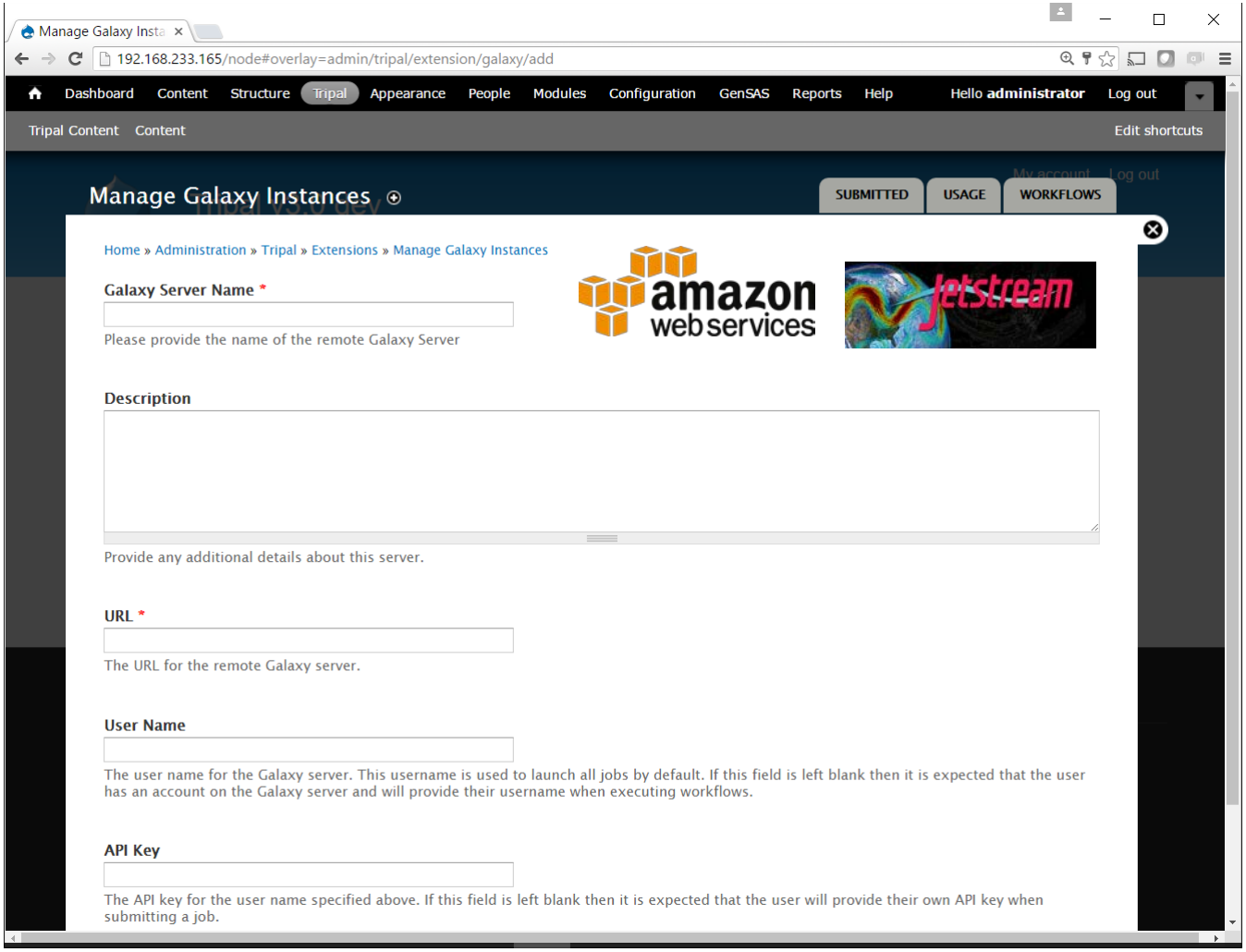
Galaxy Integration

blend4*php*

- 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!



Integrating Galaxy with Tripal

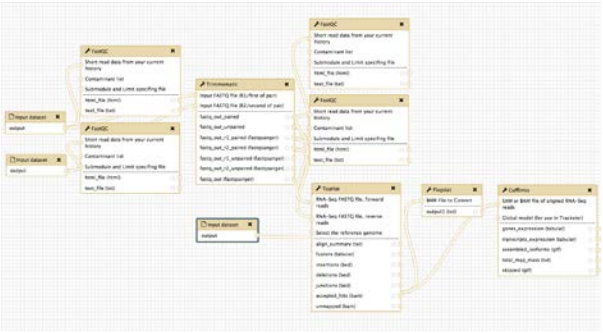


Galaxy Workflows

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**

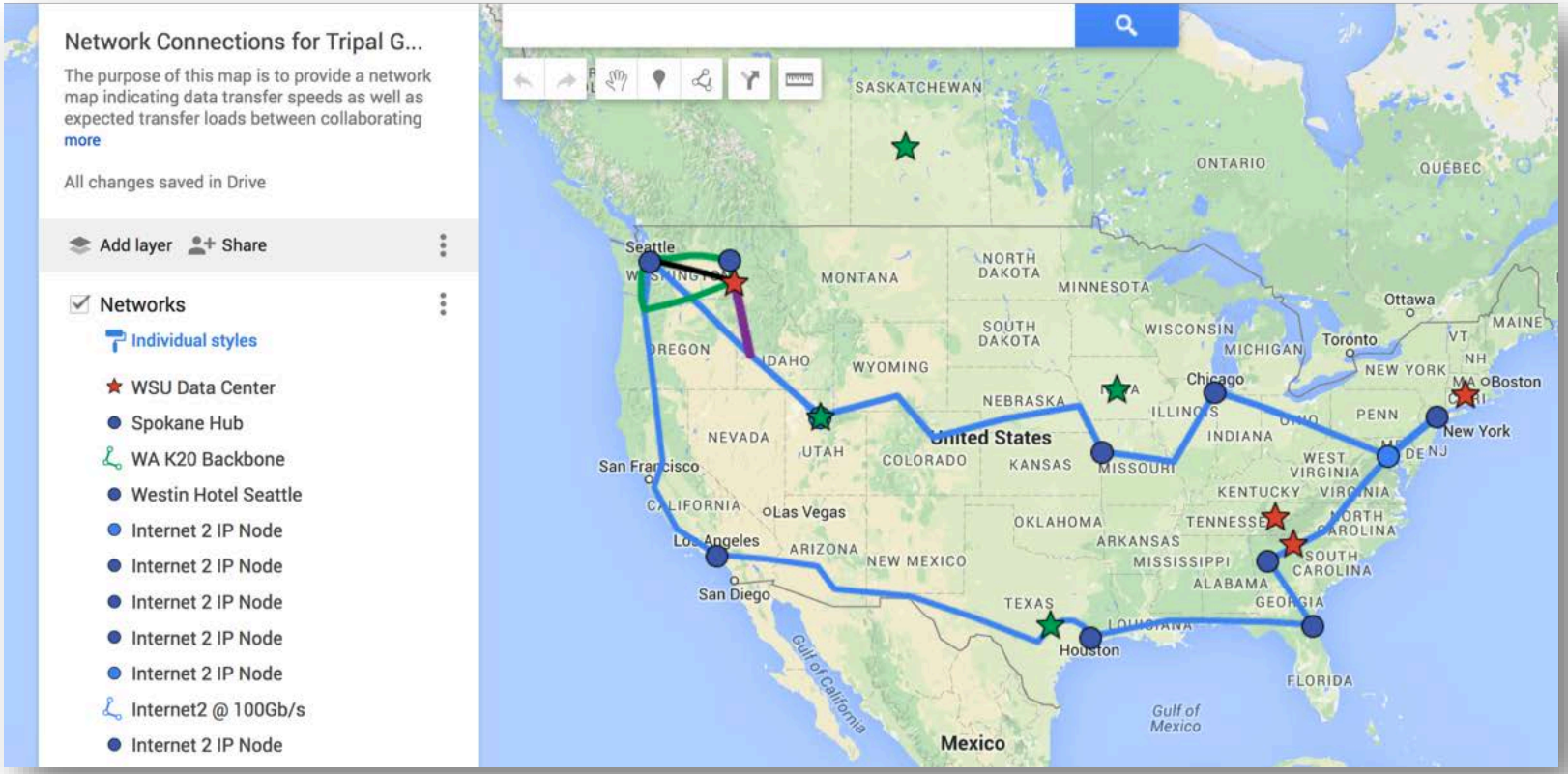


RNA Sequence Data

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

TreeGenes Database:
Software Defined Networking

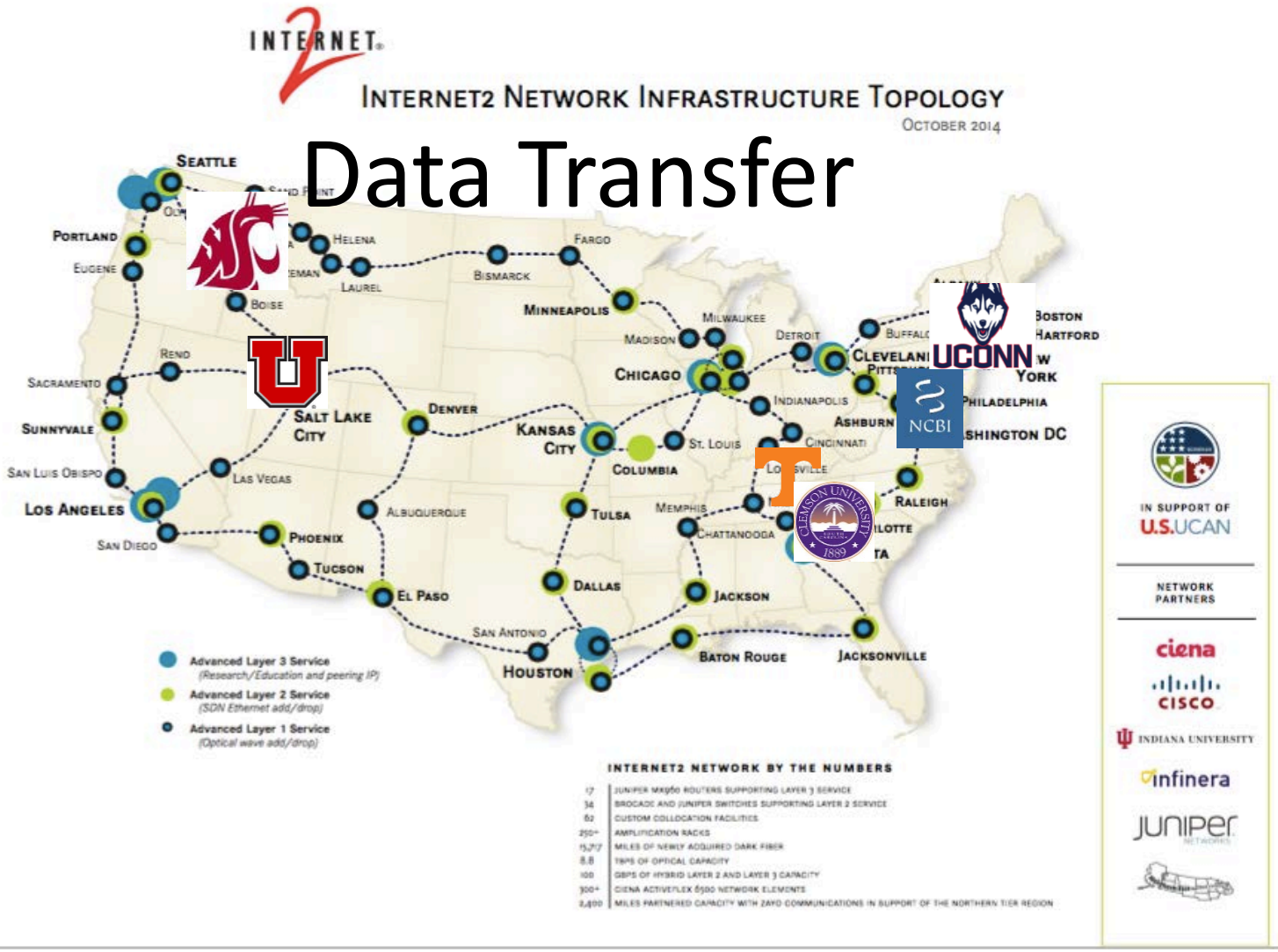
treegenesdb.org



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

Data Transfer



Tripal Gateway Use Cases

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

Tripal Gateway:

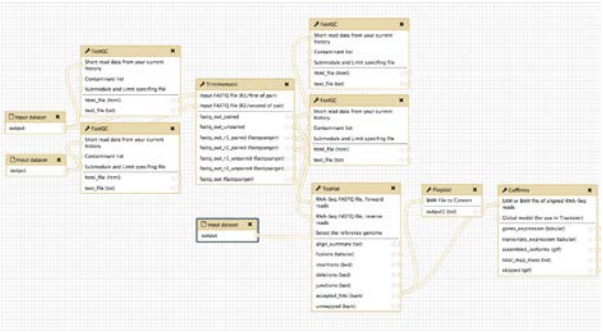
1. A user could search across community DBs for their gene of interest (by BLAST or by functional annotation keyword) using Tripal Exchange.
2. The sequences could be gathered as a list and transferred to the user with the Data Transfer (BDSS) tool.
3. If the user prefers to use Galaxy for analysis, 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

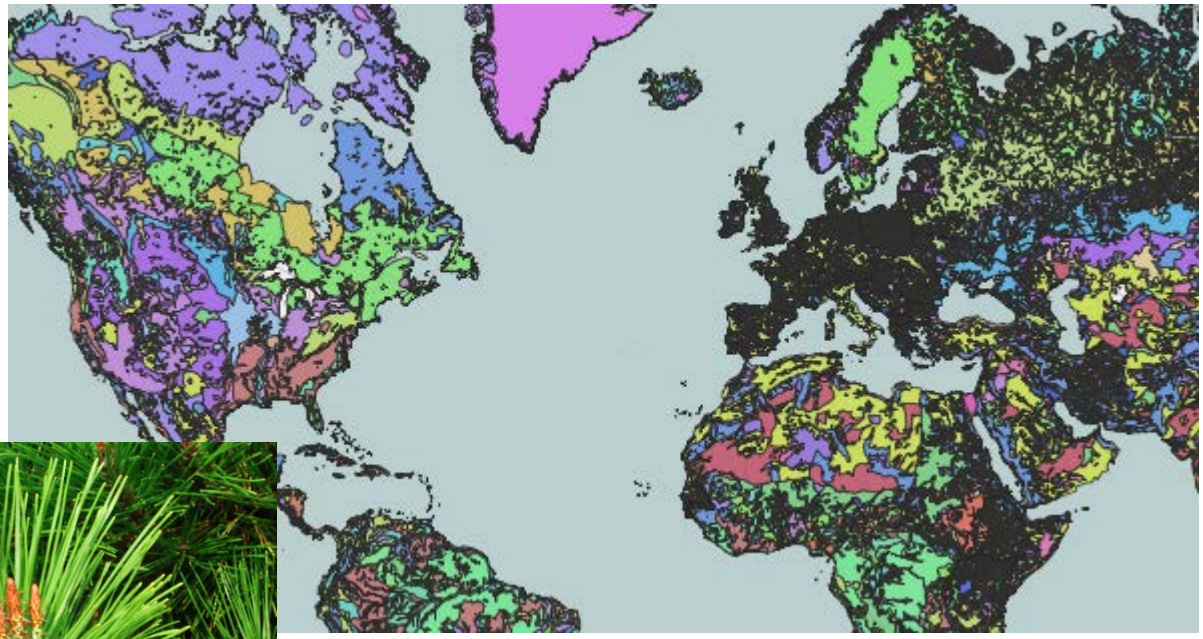
- Re-sequencing alignment
- Variant discovery (against the reference)
- **Variant discovery (between samples)**
- Prediction of functional genetic variants
- **Association Genetics**
- **Functional Annotation**



RNA Sequence Data

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

Association mapping



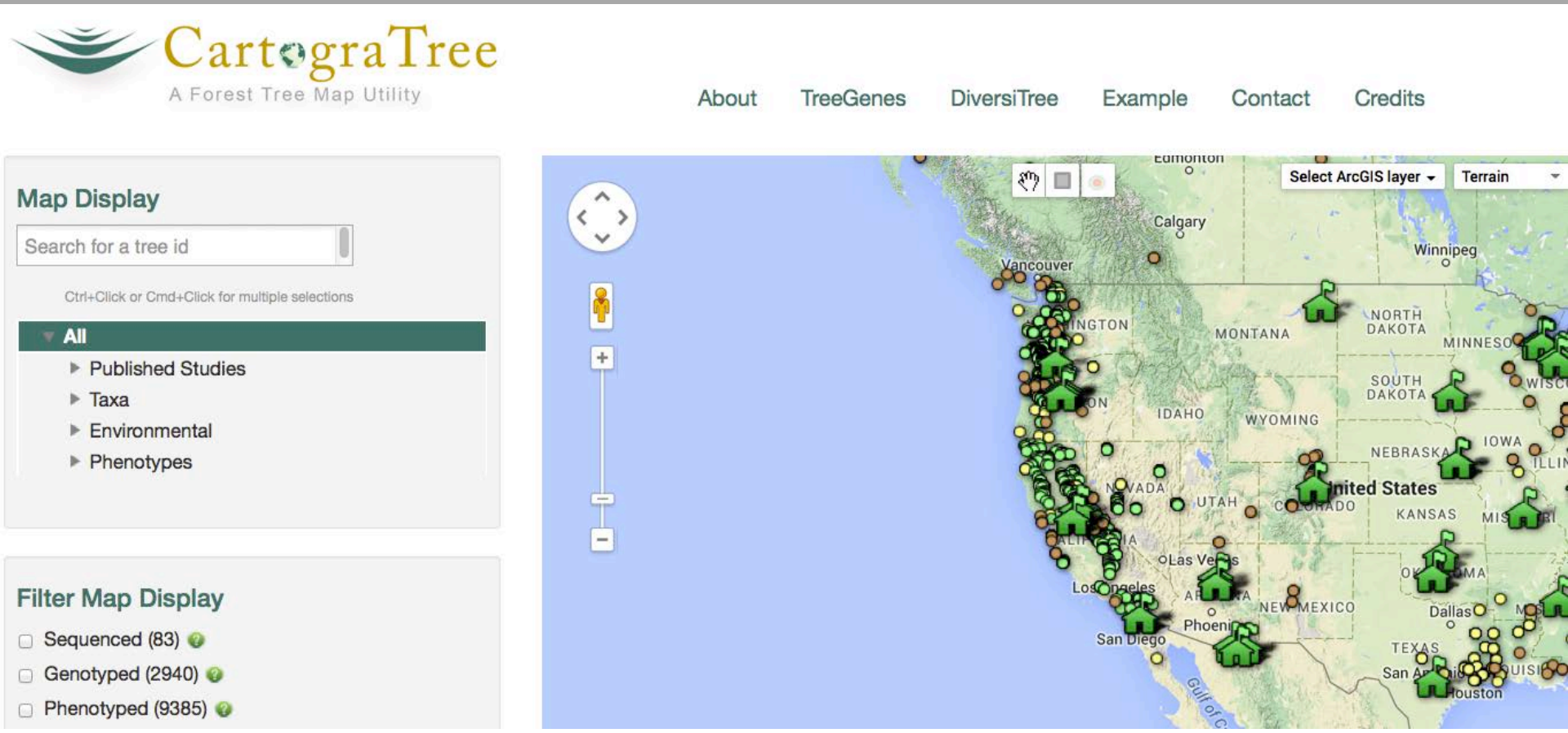
Drought and pests/pathogens changing the landscape

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

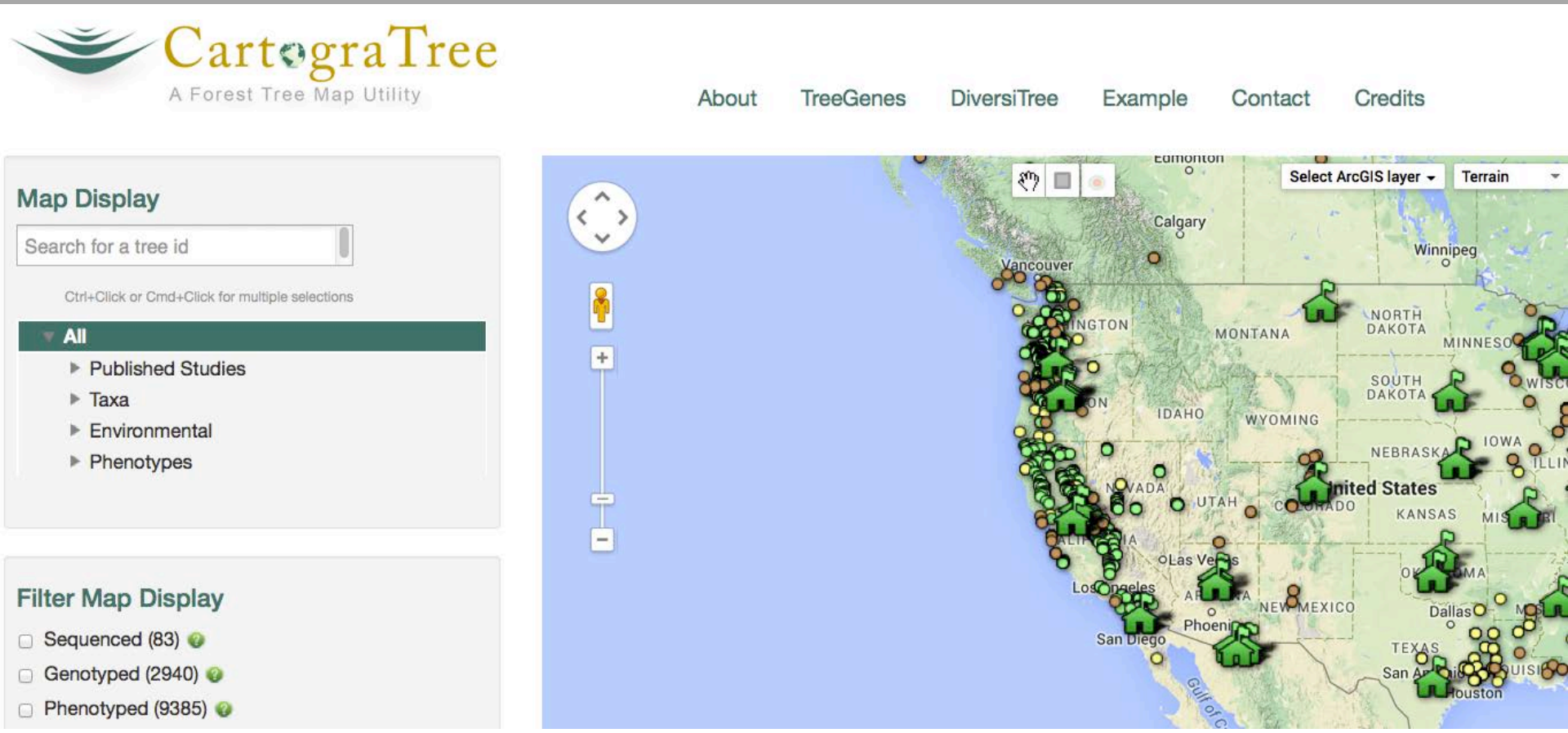
treegenesdb.org



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

TreeGenes Database: CartograTree

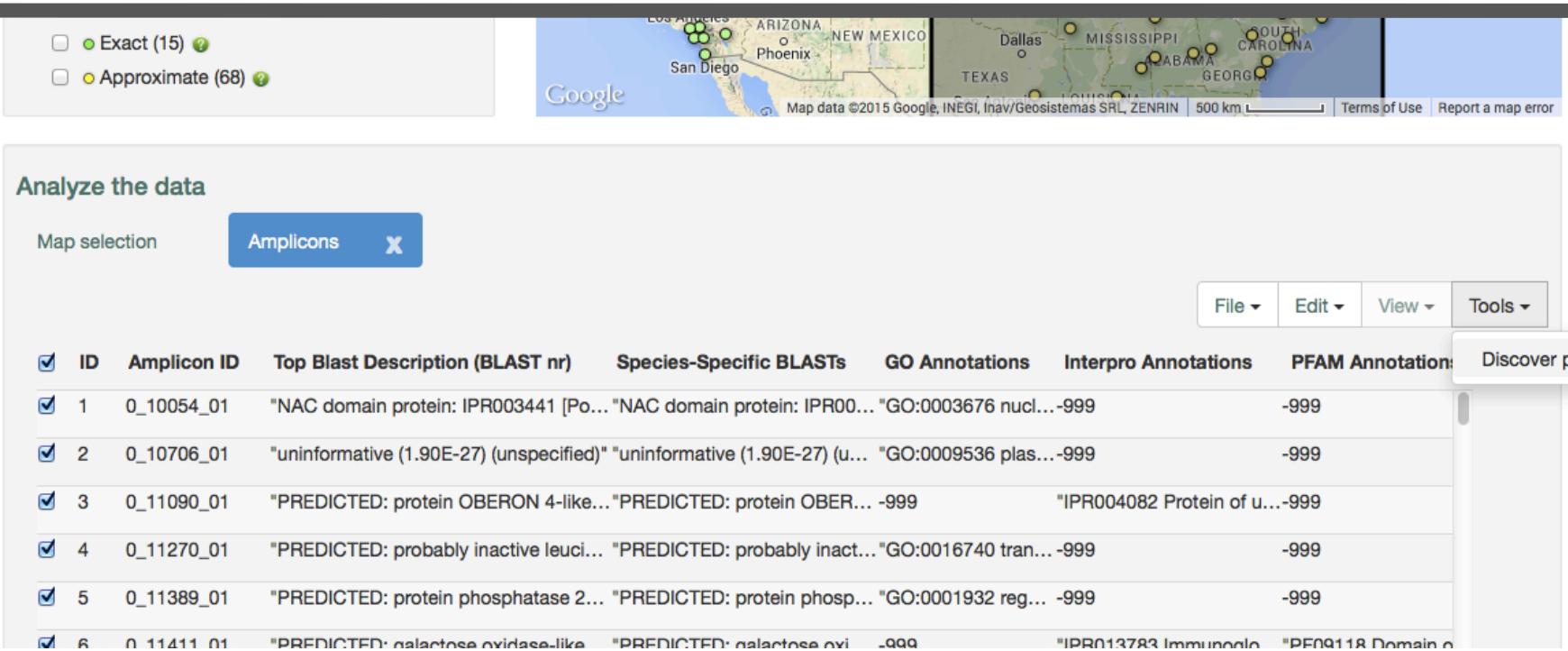
treegenesdb.org



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

TreeGenes Database: Interfaces

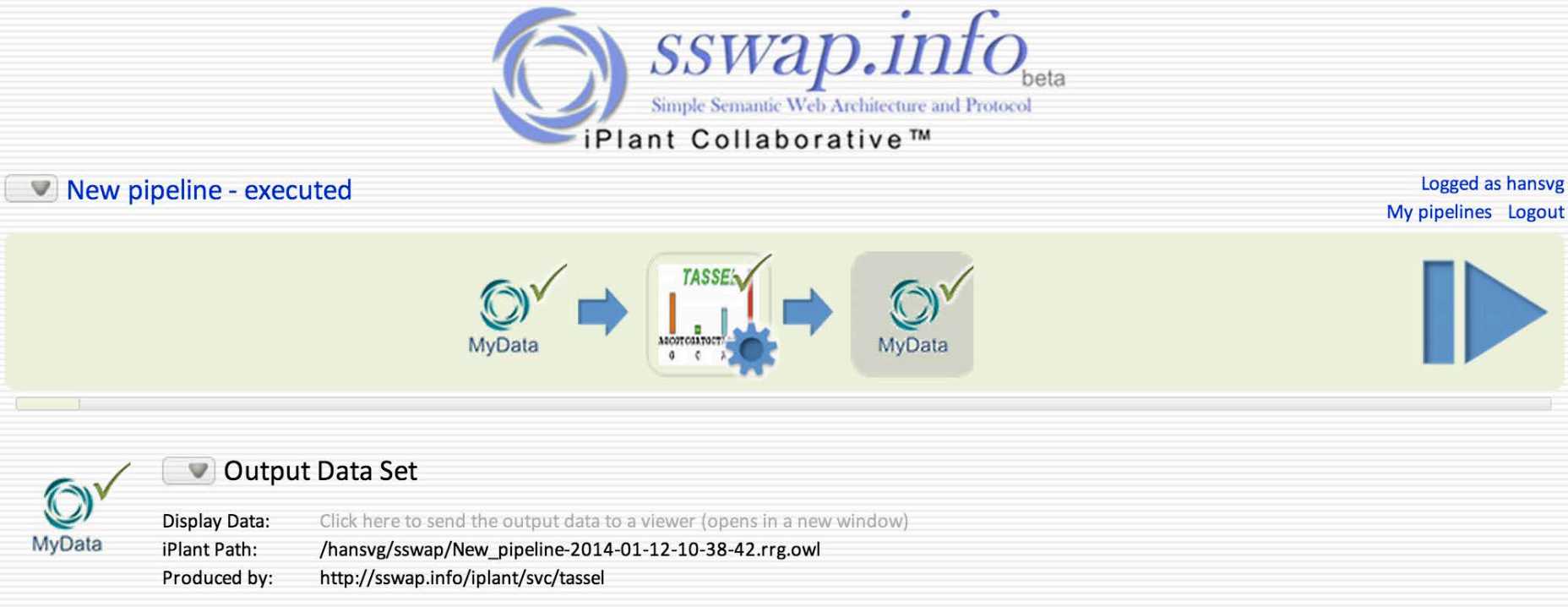
treegenesdb.org



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

TreeGenes Database: SSWAP

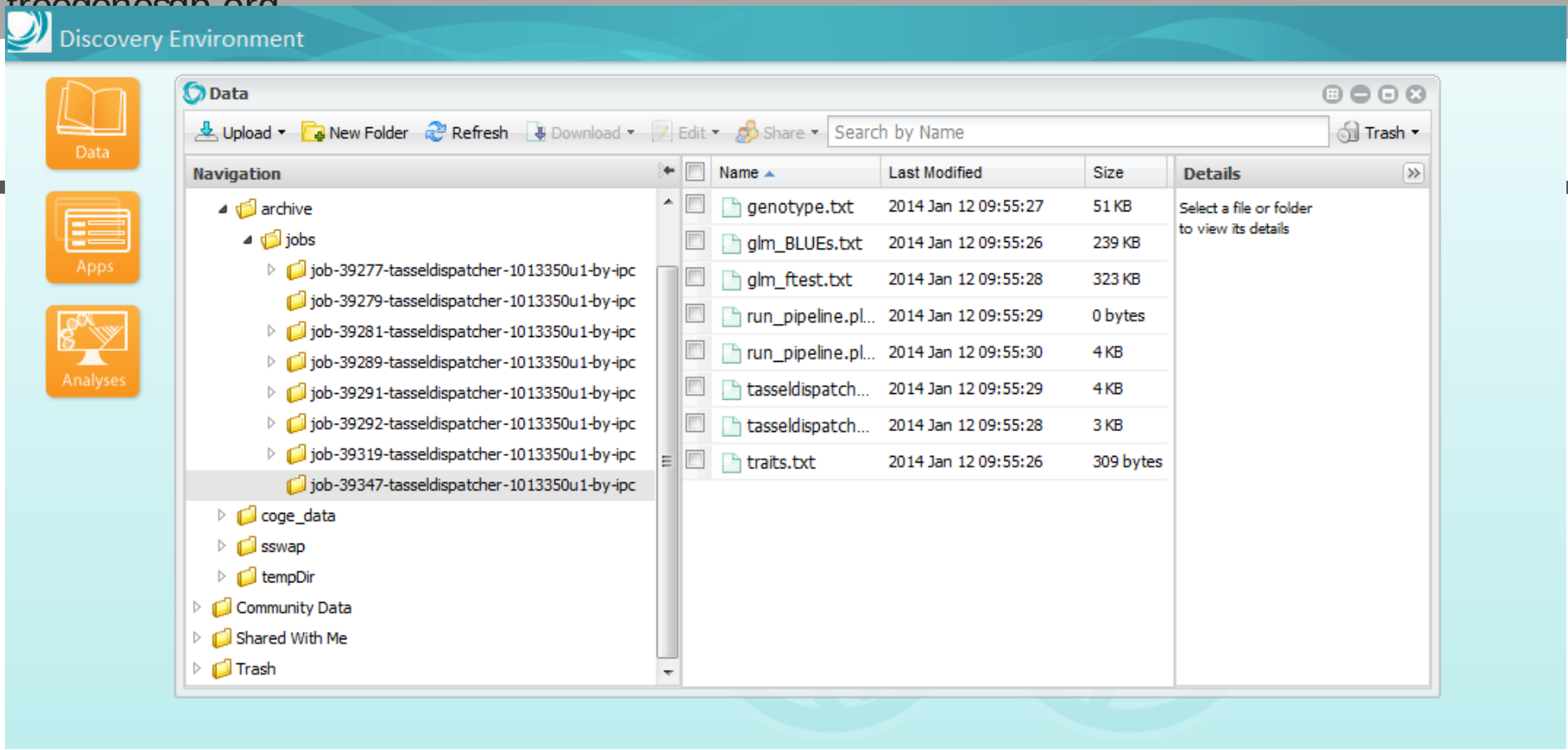
treegenesdb.org



- 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

Metadata Needed! Data Integration

TreeGenes Data Repository

TreeGenes

- **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**.

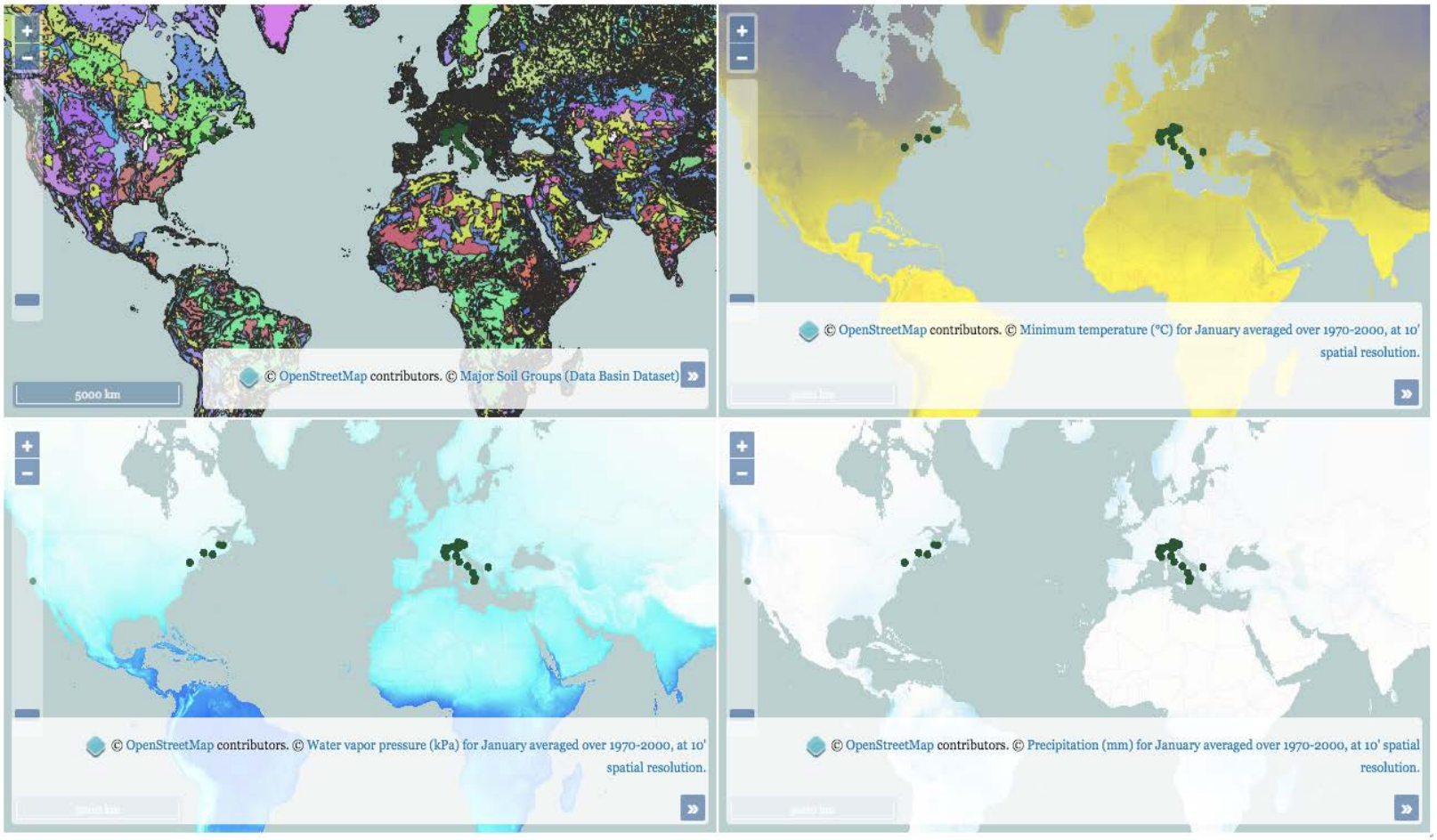
Date	Accession	Paper Title	Species	Data Statistics	Data Files
8/5/2011	TGDR001	**Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem.**	**Populus trichocarpa**	Total Sites: 1 Total Samples: 480 Total Genotypes: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

Association mapping with CartograTree

AMERIFLUX

TRY

Association mapping with CartograTree

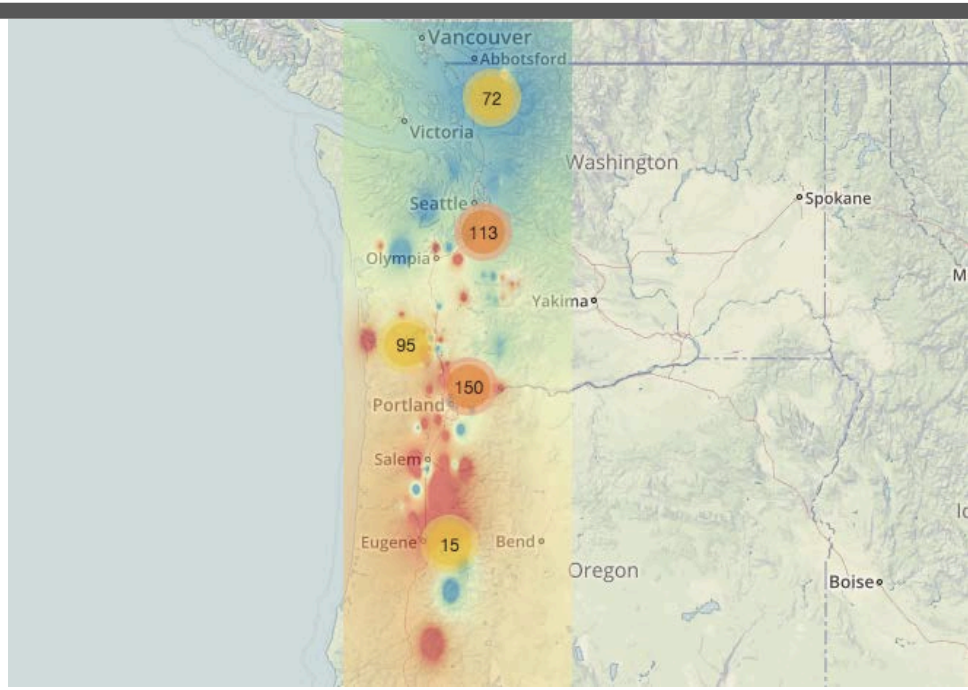


TreeGenes Database: Interfaces

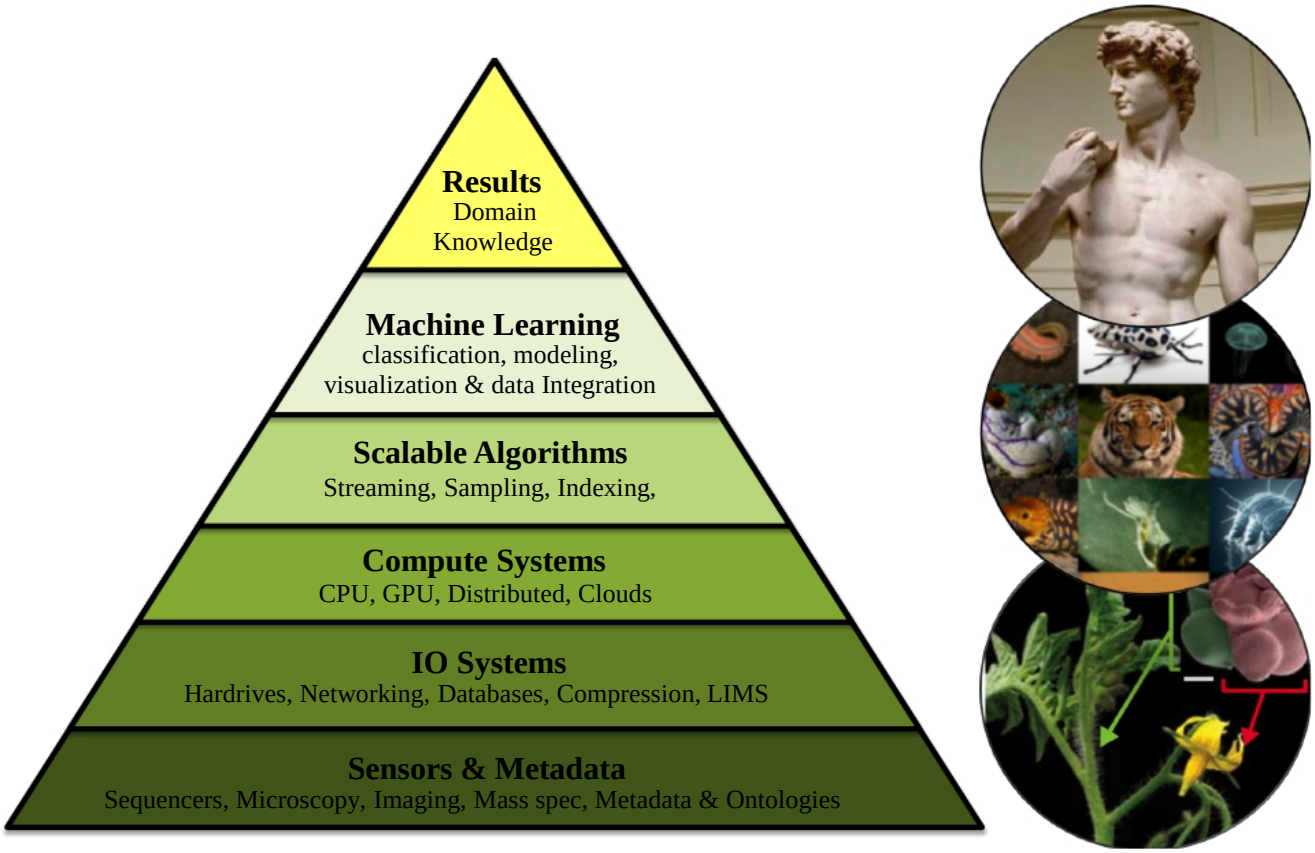
treegenesdb.org

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 it's adaptive potential and probability of survival in diverse and changing environments.

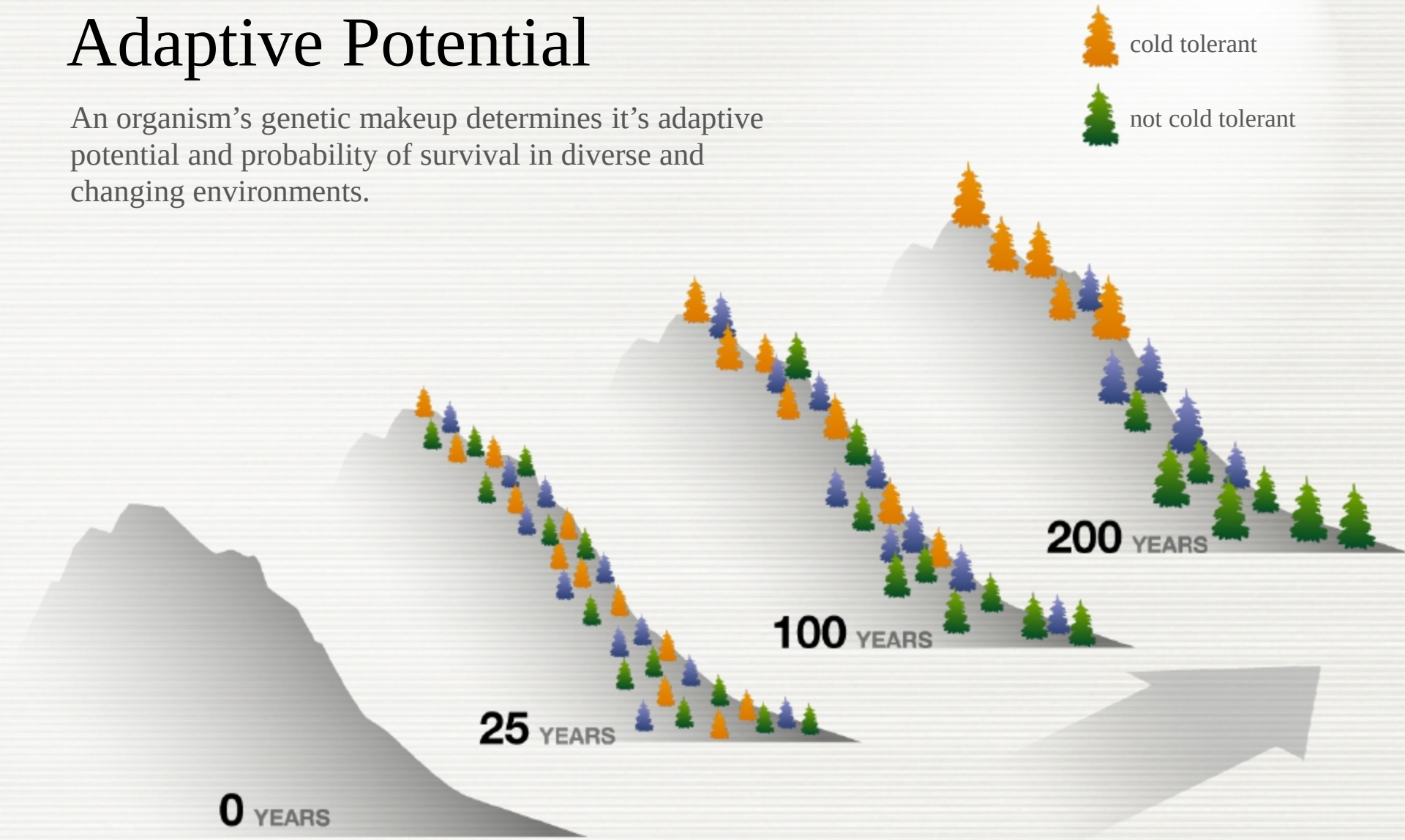


Figure credit: Nicholas Wheeler, University of California, Davis

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

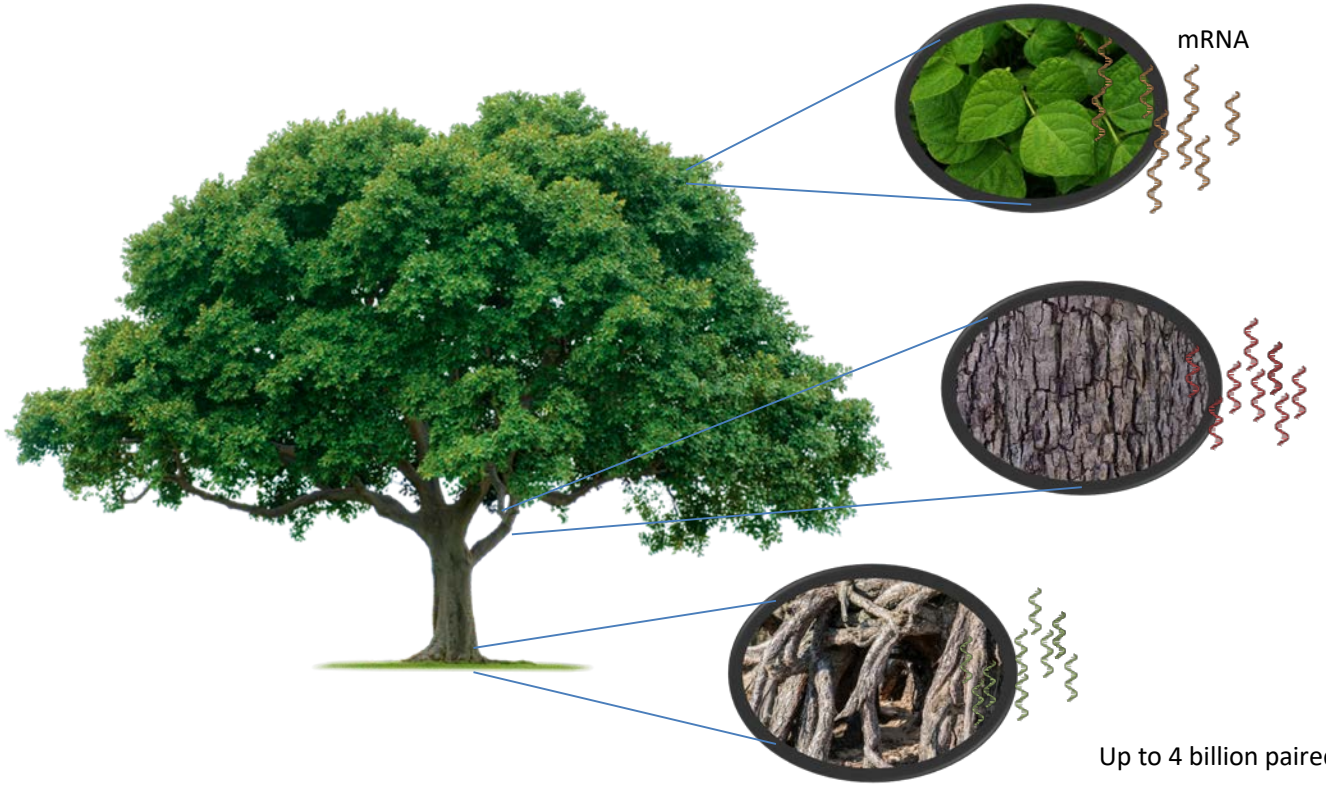
Association Genetics

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

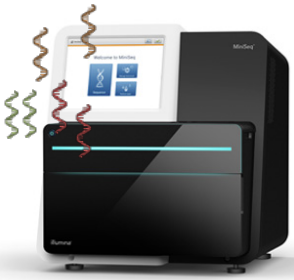
Improving Genomes

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



mRNA

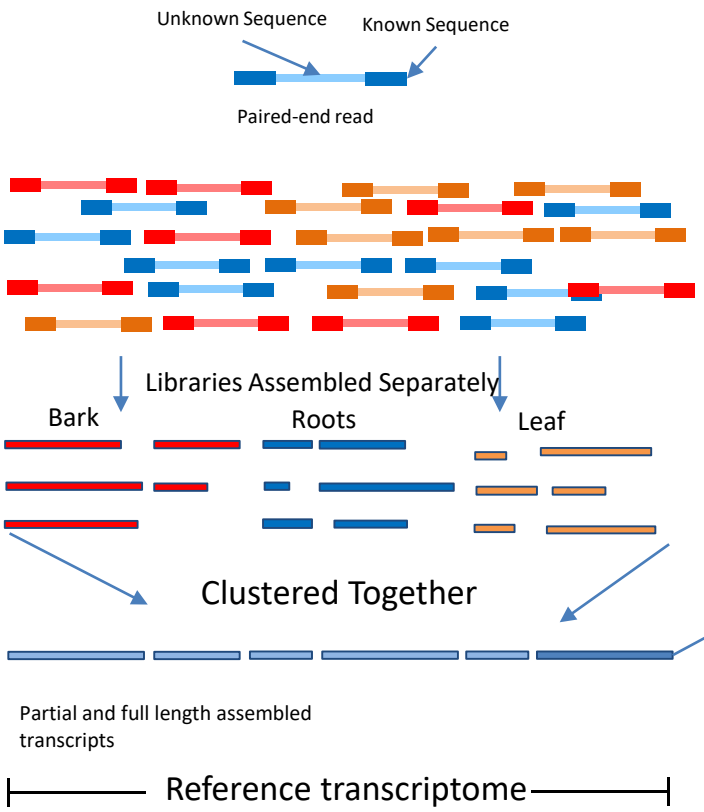


Illumina Sequencer

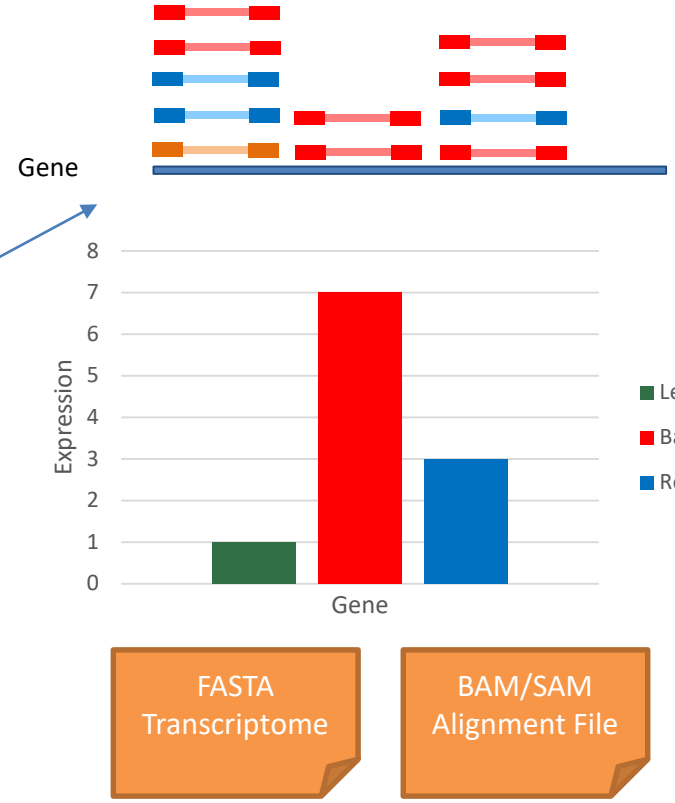
Reads

Up to 4 billion paired-end reads from a single flow cell (HiSeq 2500)

Transcriptome Assembly



Mapping to Assembled Genes



EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline

FASTA Transcriptome

BAM/SAM Alignment File

- **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.

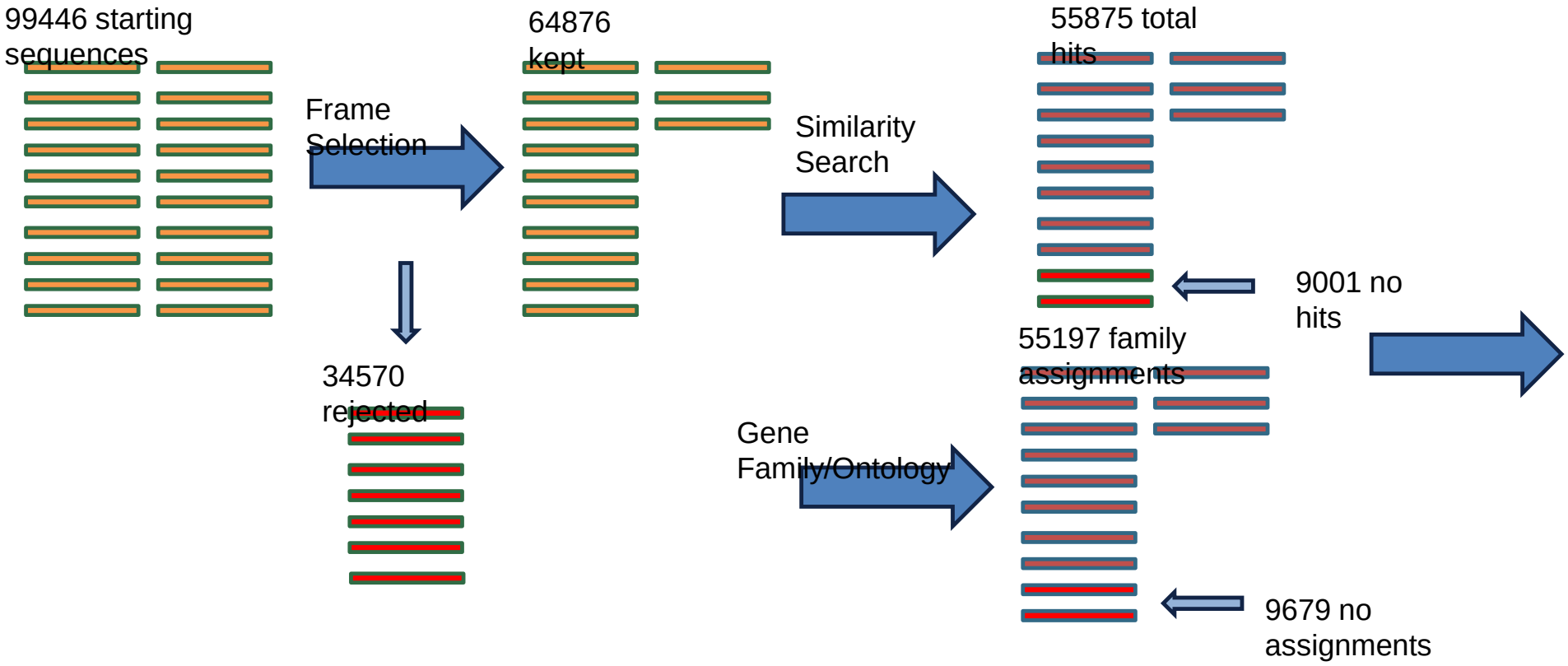
Species	Genome Size (Mbp)	N50 (bp)
J. regia	668	464,955
H. vitripennis	2200	776,706
P. menziessi	14500	387,073

EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline

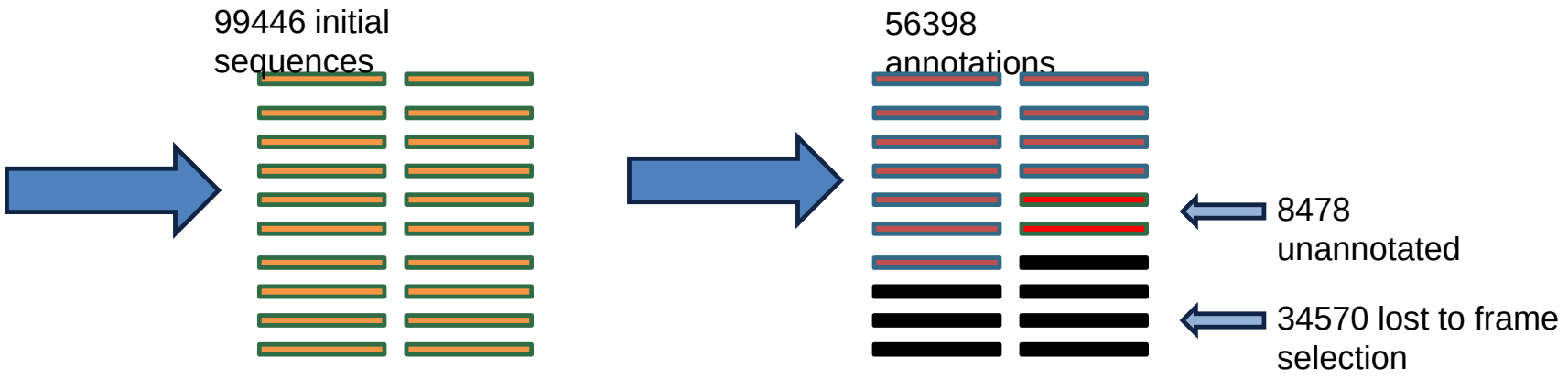
- 100,000 sequences
  - Frame selection
  - Similarity Search
    - Uniprot Swiss-Prot
    - NCBI Refseq Complete
    - Arabidopsis
  - Eggnog

- 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

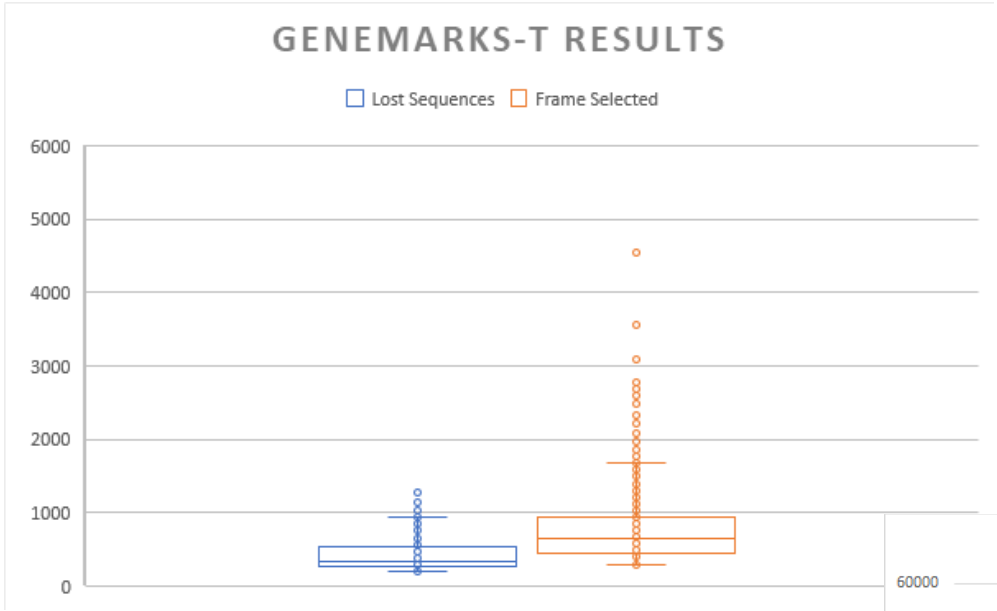


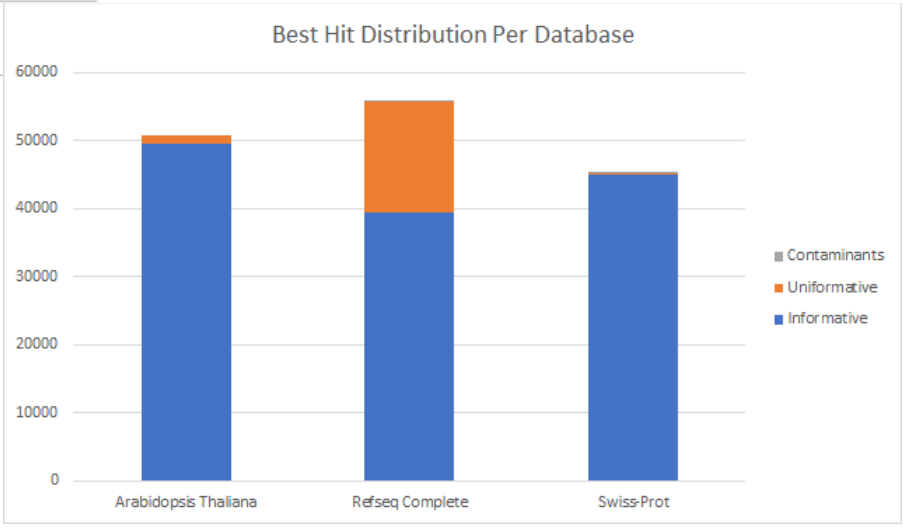
EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline



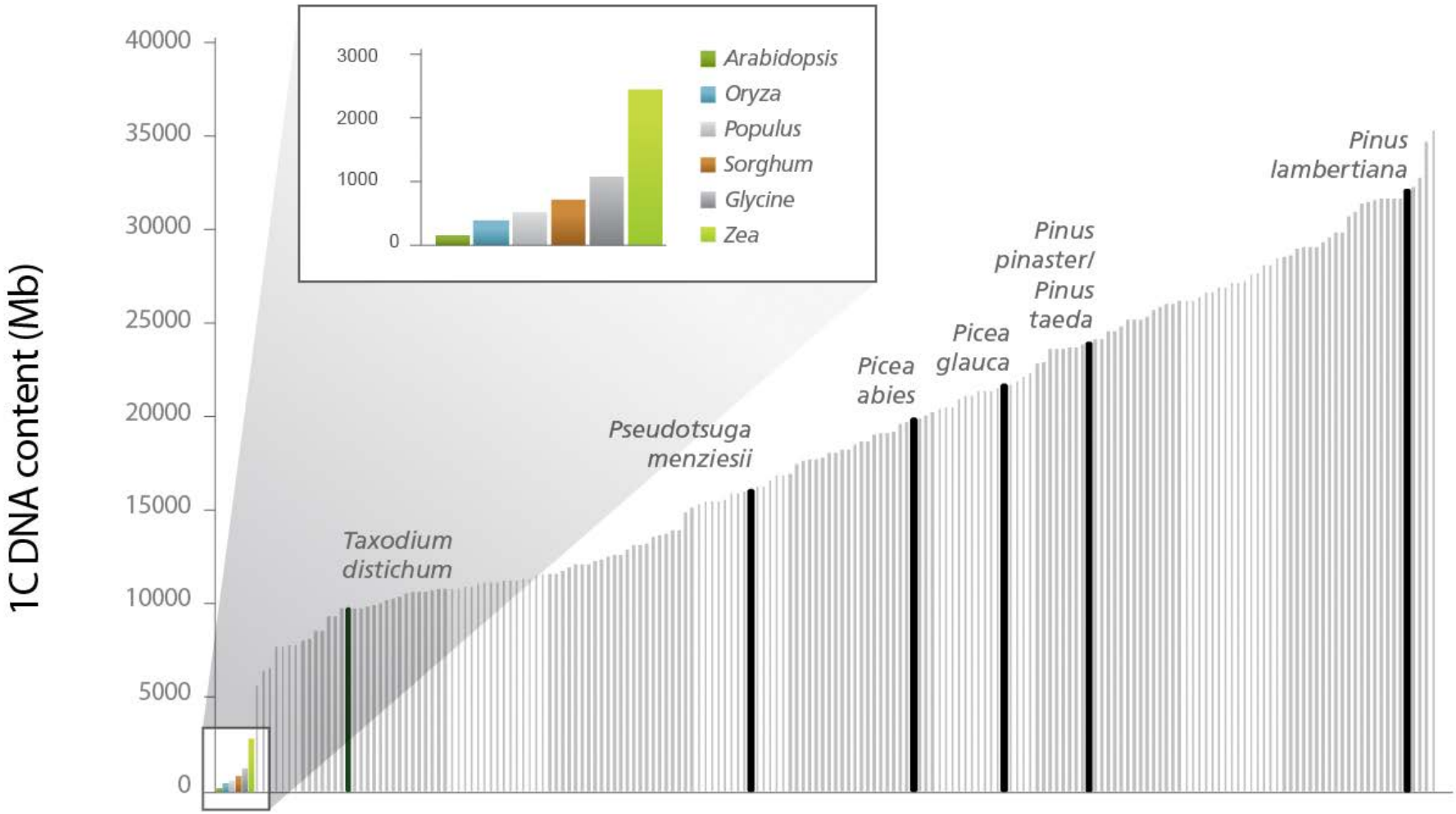
- 539 (only similarity search annotation)
- 523 (only eggnog annotation)

EnTAP: Eukaryotic Non-model Transcriptome Annotation Pipeline





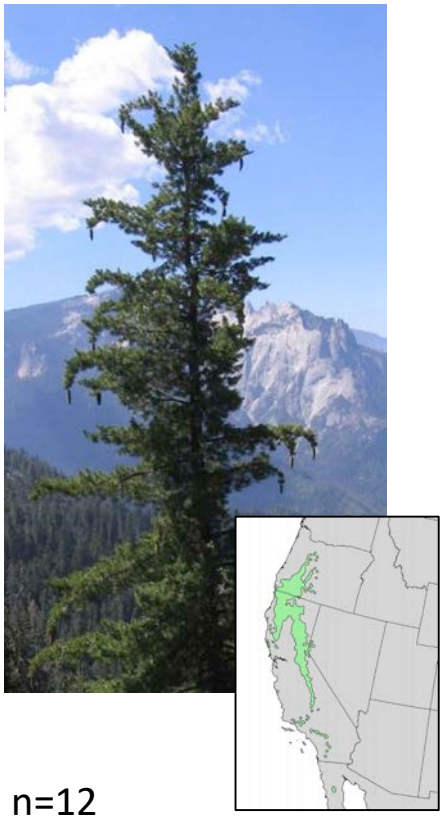
PERSPECTIVE: CONIFER GENOMES



Loblolly pine
(*Pinus taeda)*

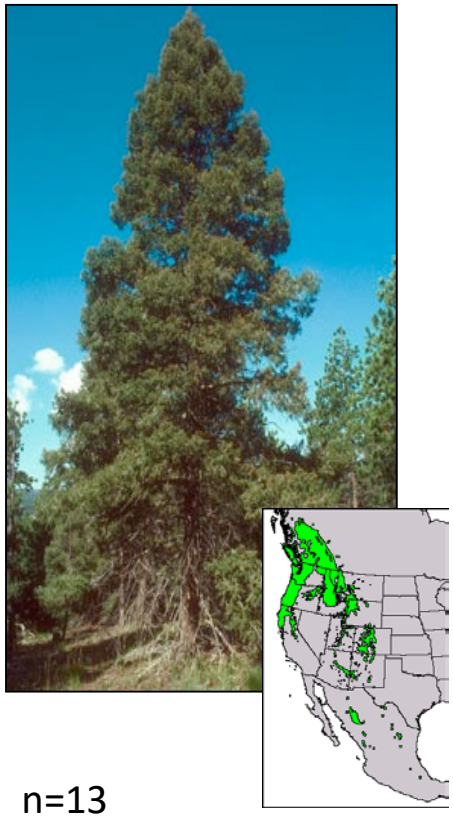
- n=12
- Genome size: 21.6 Gbp
- Genotype to sequence: 20-1010
- Mapping population: 6-1030x8-1070 and 20-1010x11-1060 (1000 F_1 progeny)

Sugar pine
(*Pinus lambertiana)*



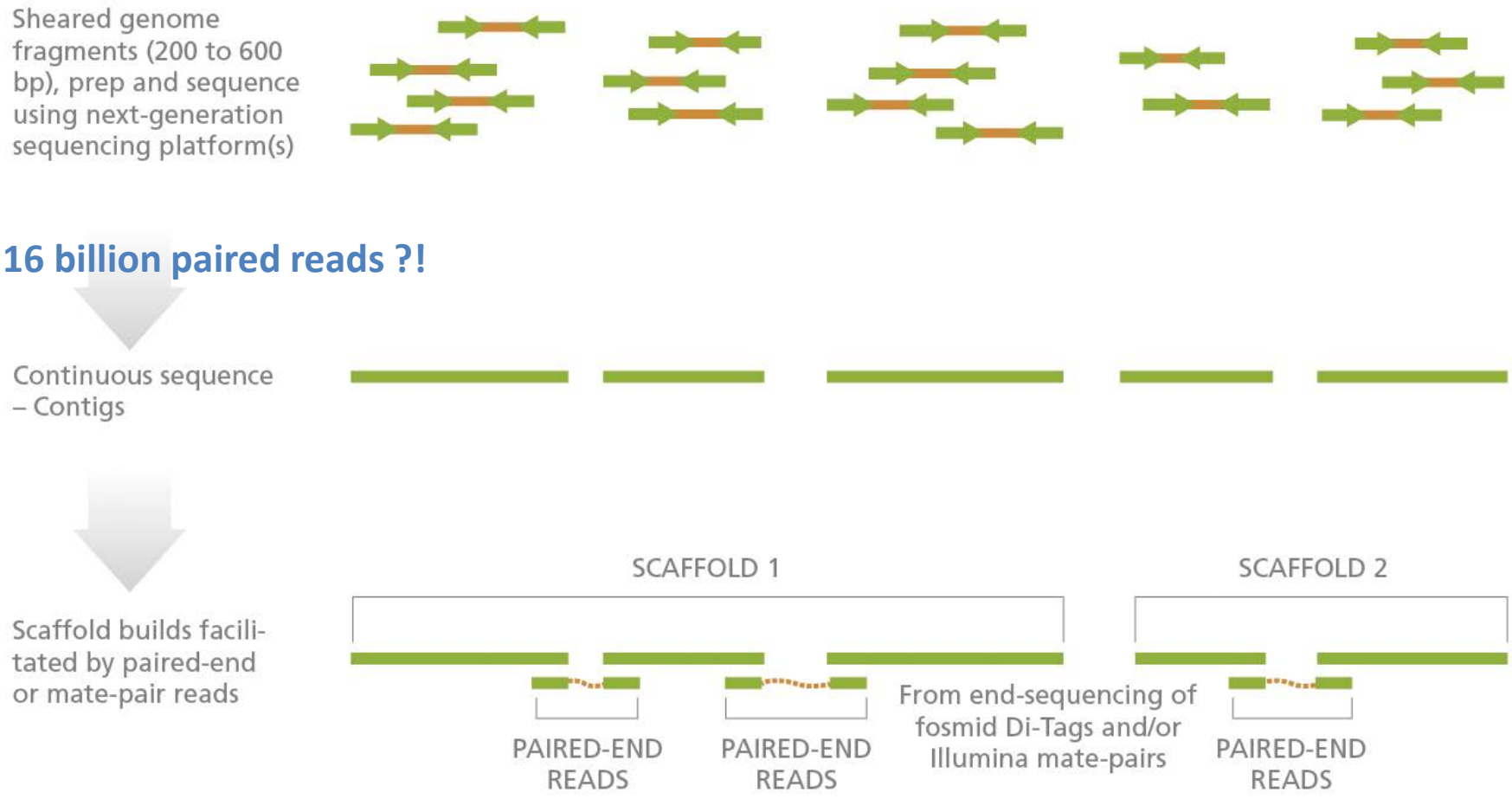
- n=12
- Genome size: 31.9 Gbp
- Genotype to sequence: 6000
- Mapping population: 5038x5500(1300 F_1 progeny)

Douglas fir
(*Pseudotsuga menziesii)*



- n=13
- Genome size: 18.6 Gbp
- Genotype to sequence: 412-2
- Mapping population: 412-2x013-1 (1000 F_1 progeny)

ASSEMBLING THE REFERENCE GENOME (WGS)

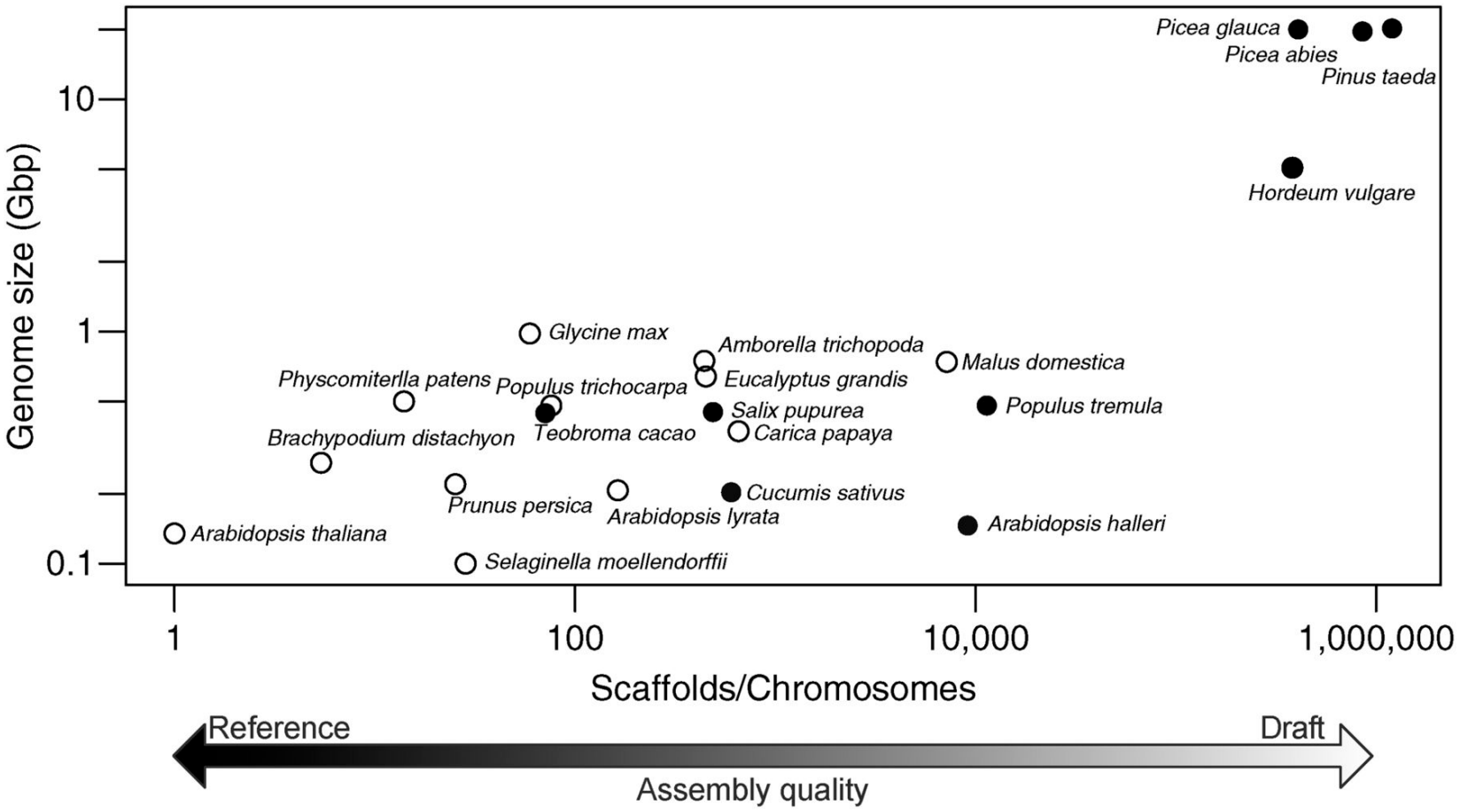


Conifer Genomes Compared

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

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

Genome Assemblies Compared



Amanda R. De La Torre et al. Plant Physiol. 2014;166:1724-1732

Genome annotation

~32 billion bp

ACAATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATT
GATACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCA
ATAAATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTT
CTAGGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGA
TCAGATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAA
GTAAATAAAGTTAATTTCAAGTTGTAATTGATGCTAGAAAGACAATGAAAC
AGAGCCATGTGACCAATGAGAGAGATGAGGGTGGCAGCAGCCTGTTTT
AGATAAGGTACCTGATTGGTGGGATTGGAAGACCTCTCTGAGATTAGTG
TCTTCAGATATGCCTTAATGATATGAAAGAACCATTCATGGGAAGGCCTA
GCATTAAAAACCGTCTAGGCAGAATGAGCAGCAAGTGCAAGGGTCCTG
GATAGGAATGAGCTGGATATACTCAAGGAAGAAAGAGAAACTATGGAAA
AATGAAAATAGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTT
ACTTTTCTTCTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAAC
AATAAATCACATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGA
TACCTTTAAATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATA
AATATTTTTTAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTA
GGCATTGGGGATACCATGTTCACAAGACAGACTATGATTTACAGGATCA
GATGTGGACTCTCAAATTCGACTGAGAATAAAACAGACACTAAACAAGTA
AATAAAGTTAATTTCAAGTTGTAATTGATGCTACTATGGAAAAATGAAAAT
AGATTTTAAAACATGTTAATTCACGTTACTTTTTGTTAAATTTACTTTTCTT
CTTTCACTTCTTACCTGTCAATGTTATTAATATTTTTAGGAACAATAAATCA
CATTAATTCCTTATCTCATGTGAAATTTCATATTTATGATTGATACCTTTAA
ATGTCATTTGTTGAAGGAAGATTATTCATTTTTTCATTCAATAAATATTTTT
TAGAATAATAAGTCCCAGGCACAAGACCAGTATTATGTTCTAGGCATTGG
GGATACCATGTTCACAAGACAGACTATGATTTACAGGATCAGATGTGGA
CTCTCAAATTCGACTGAGAATAAAACAGACACAAACAAGTAAATAAAGTT
AATTTCAAGTTGTAATTGATGCTATCCCAGGCACAAGACCA....

Genes:

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

Genetic variation:

- SNPs

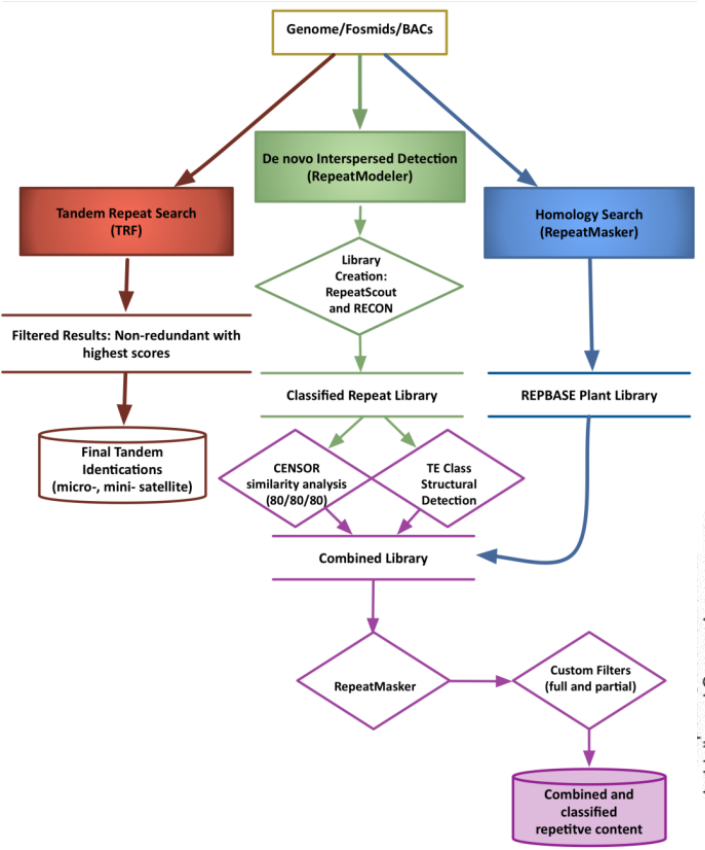
Regulatory sequences:

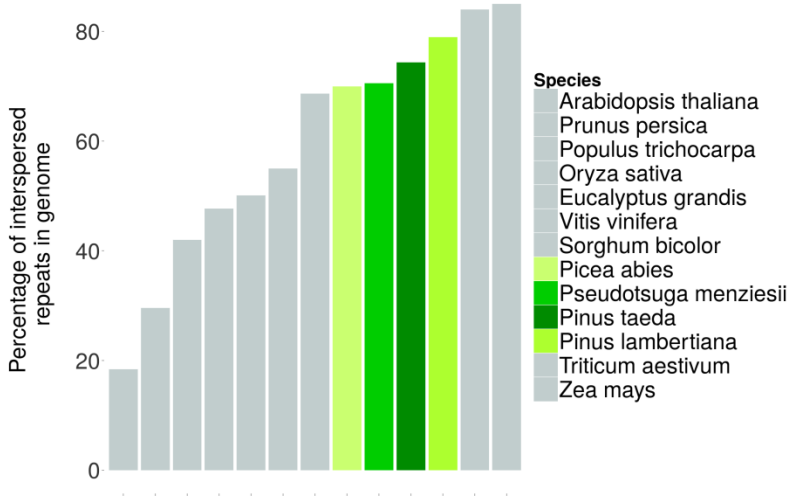
- Promoters
- Enhancers

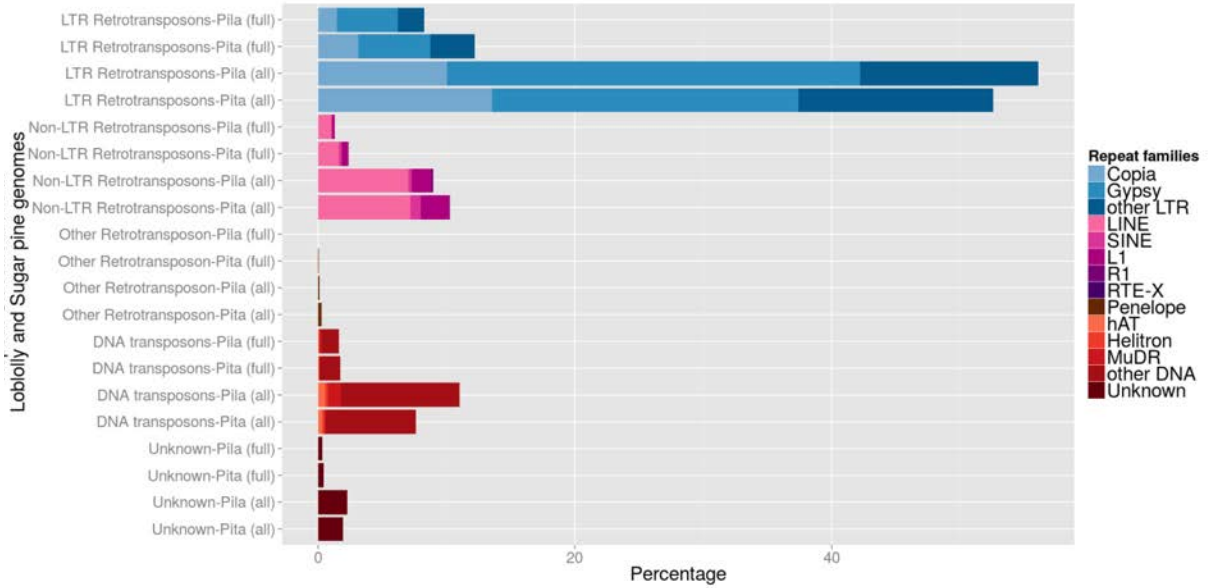
Epigenetics:

- DNA methylation
- Chromatin

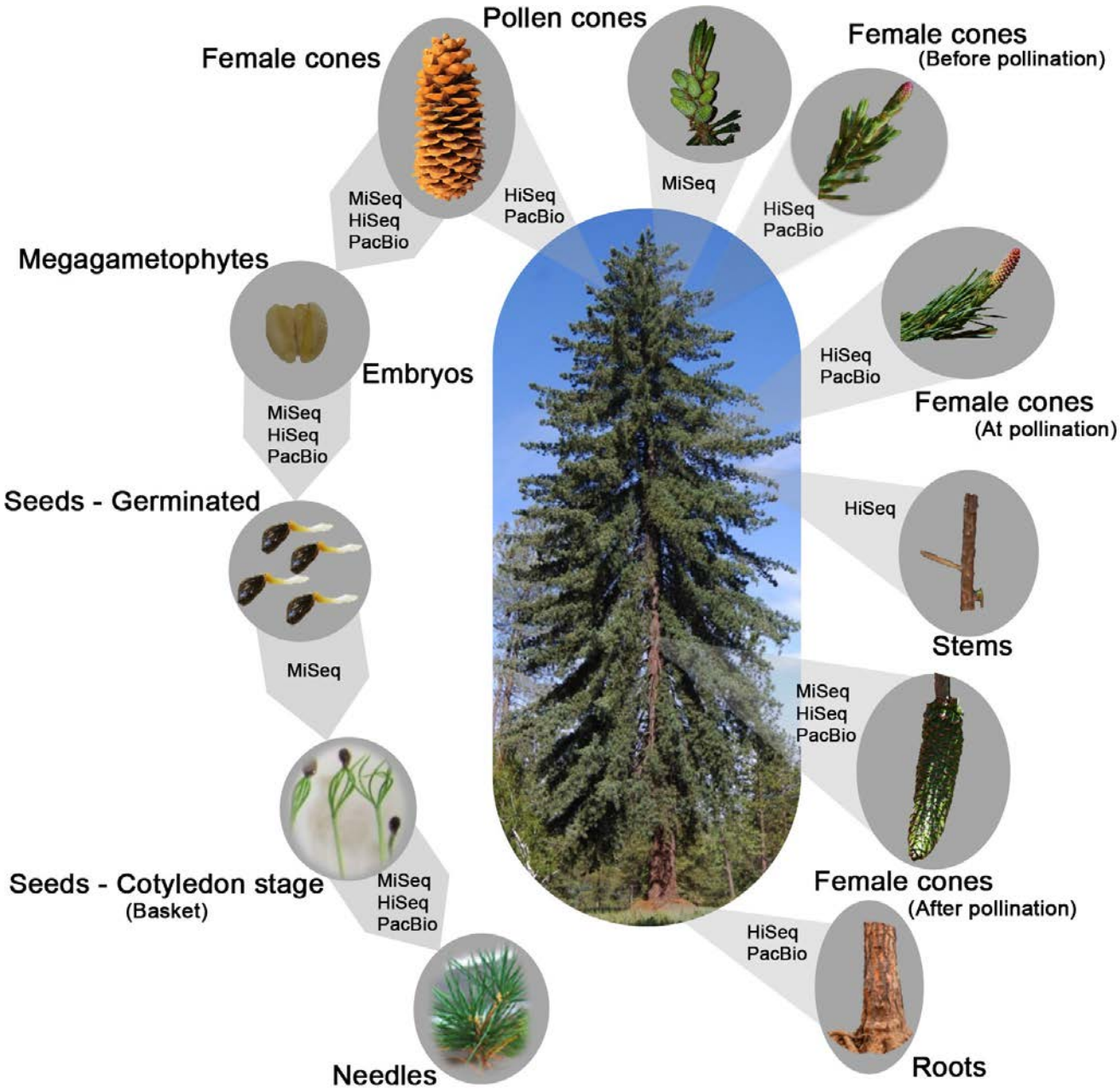
Similarity and *de novo* Repeat Identification







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


Female cones
MiSeq
HiSeq
PacBio
Pollen cones
HiSeq
PacBio
MiSeq
Female cones
(Before pollination)
HiSeq
PacBio
Megagametophytes
Embryos
MiSeq
HiSeq
PacBio
HiSeq
PacBio
Female cones
(At pollination)
Seeds - Germinated
MiSeq
HiSeq
PacBio
HiSeq
Stems
MiSeq
MiSeq
HiSeq
PacBio
Seeds - Cotyledon stage
(Basket)
MiSeq
HiSeq
PacBio
Female cones
(After pollination)
HiSeq
PacBio
Needles
Roots

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.

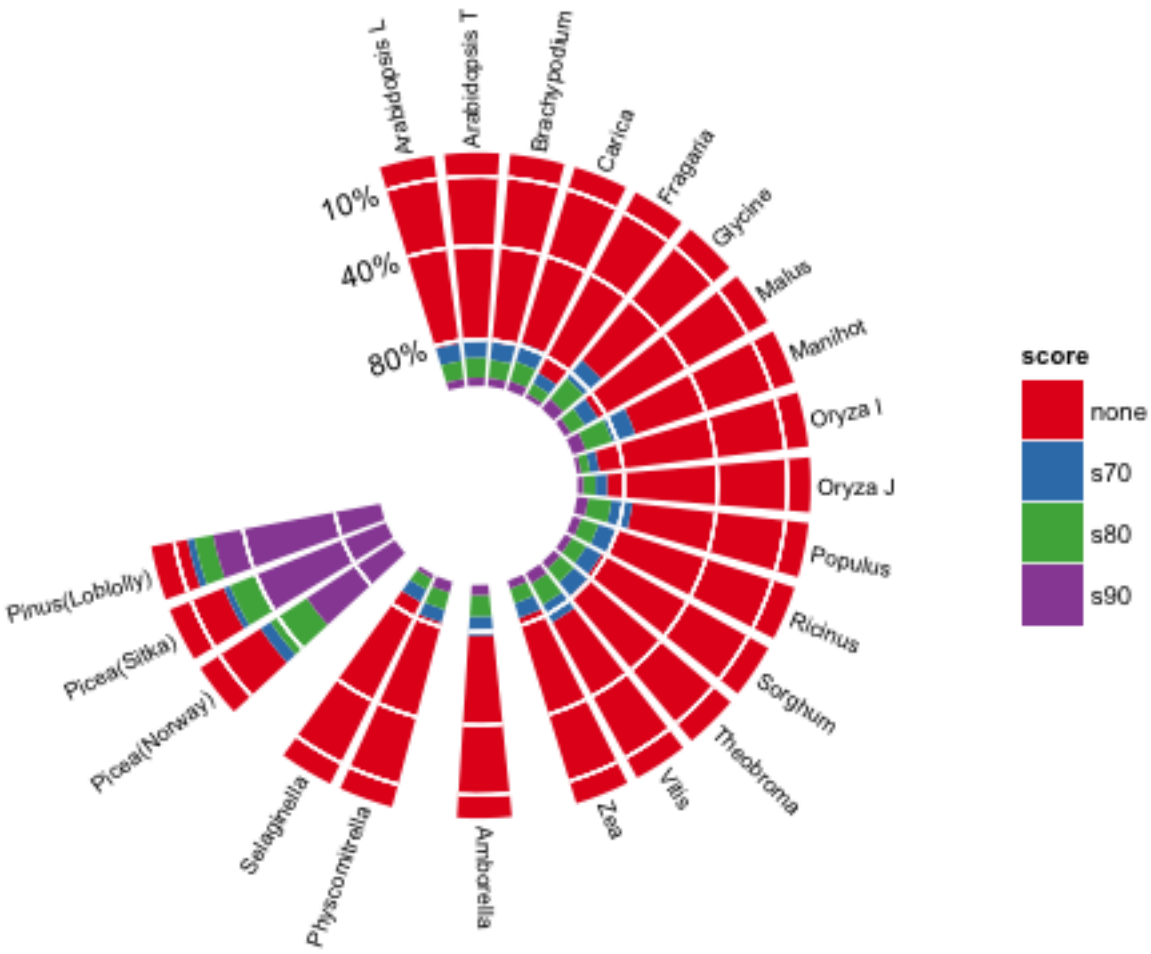
29,189 de novo transcriptome + 42,345 unique additions =

> 71,534 genes (round 1)

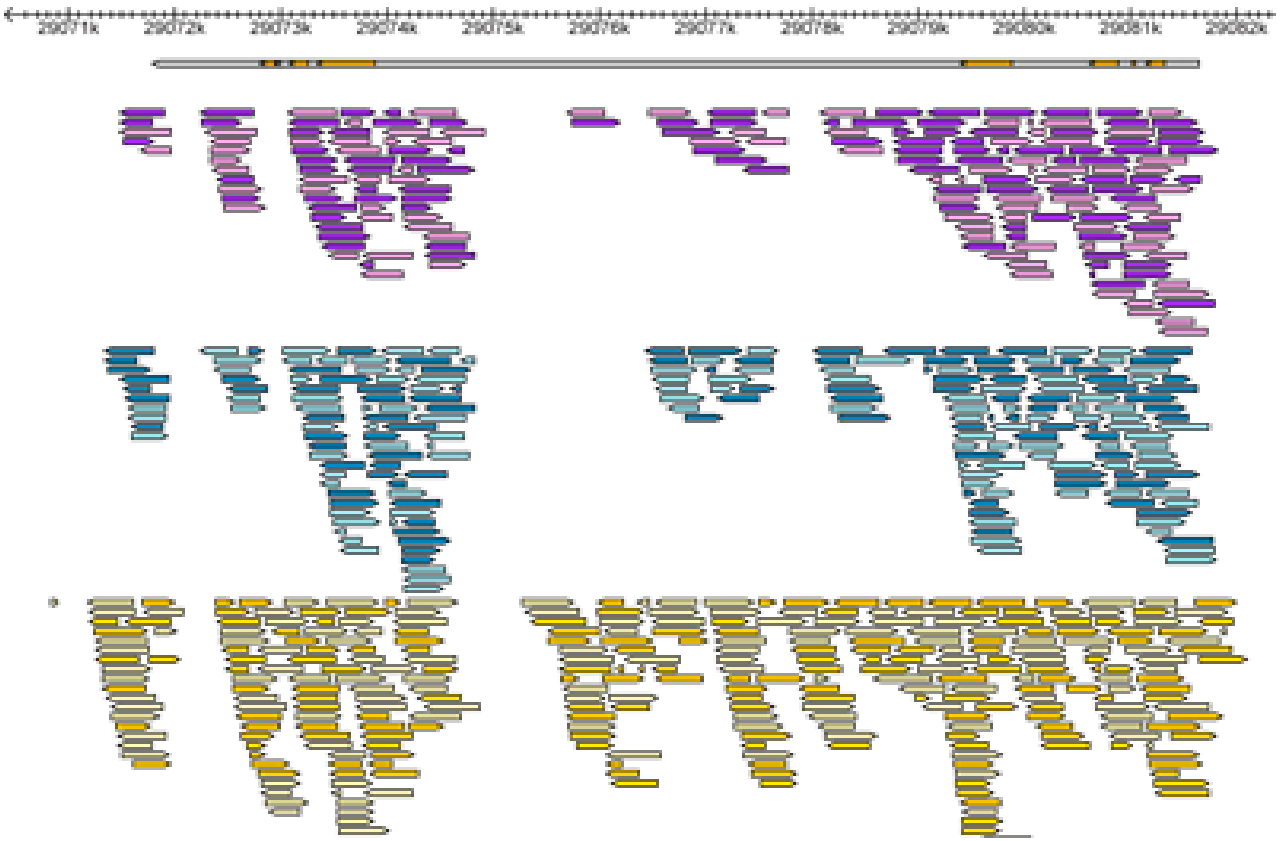
> 8,000 genes (round 2)

Challenges:

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



Improving Gene Annotation
Model developed on walnut (*Juglans*) genomes

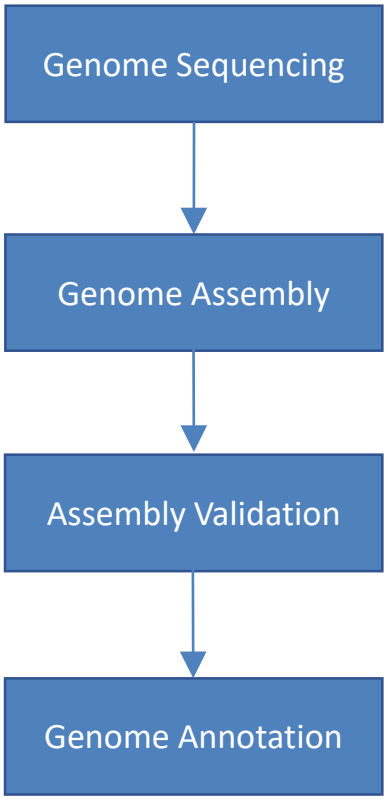


Masking + RNA-Seq Reads + Pseudogene + Assembled Evidence

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)

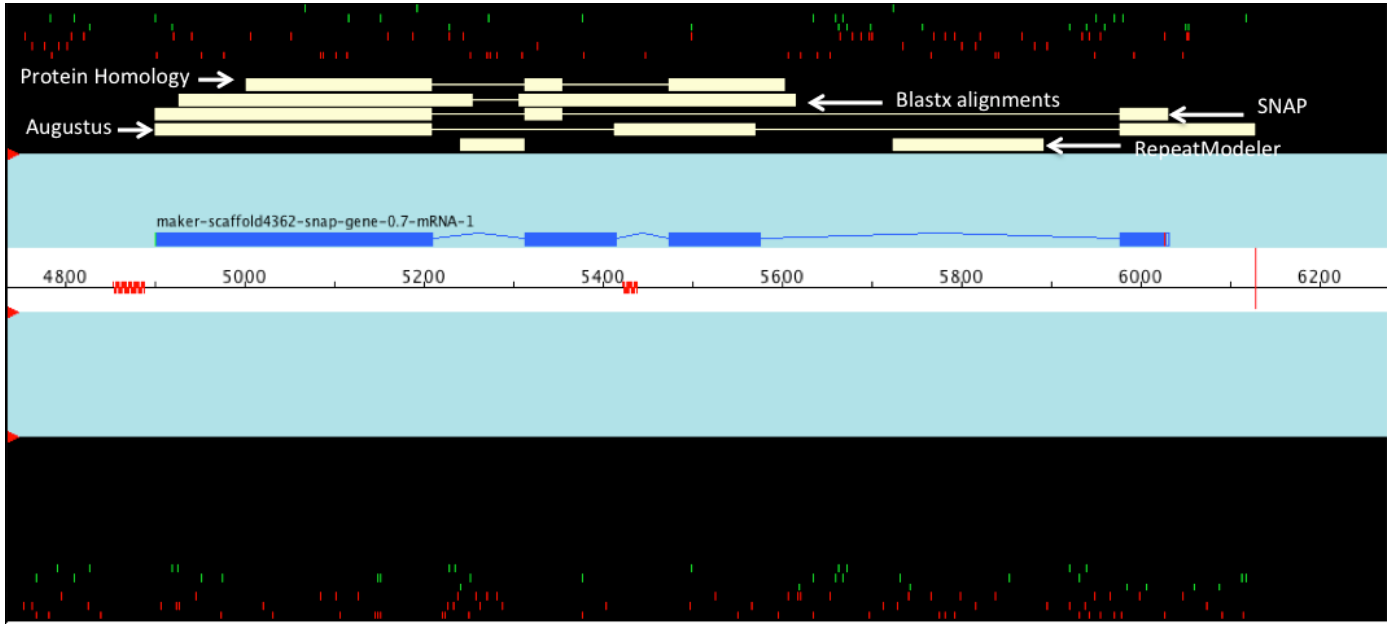
- Two different assembly methods
- Transcriptome scaffolding

- 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.



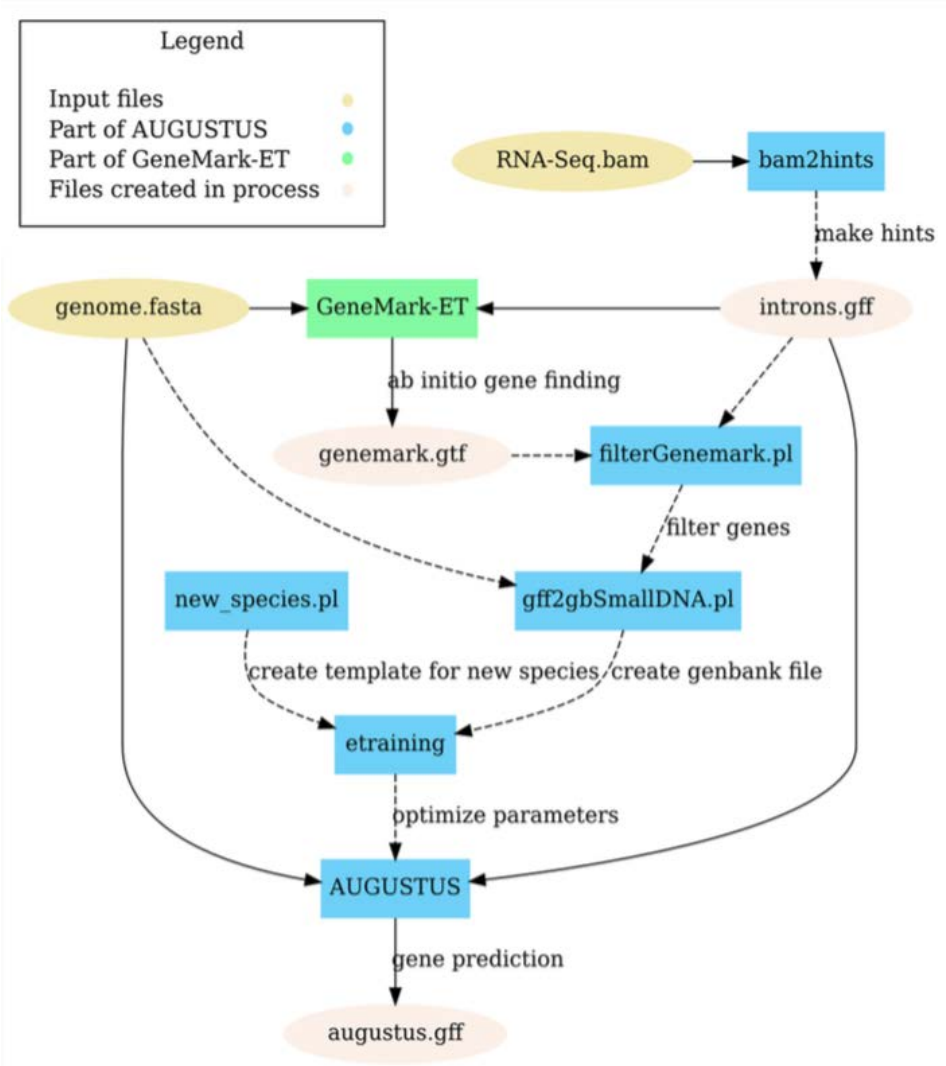
MAKER Annotation Results

- 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. GeneMarkE-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

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
Jeffry Mitton
University of California, Merced
Lara Kueppers