

Cyberinfrastructure for Landscape Genomics: Connecting Biological Databases, Metadata, and Intelligent Analytics

Jill Wegrzyn

Plant and Animal Genome Conference

January 12th 2019



Biologists need more than 'omics!



Early drivers of 'omics: Genomics

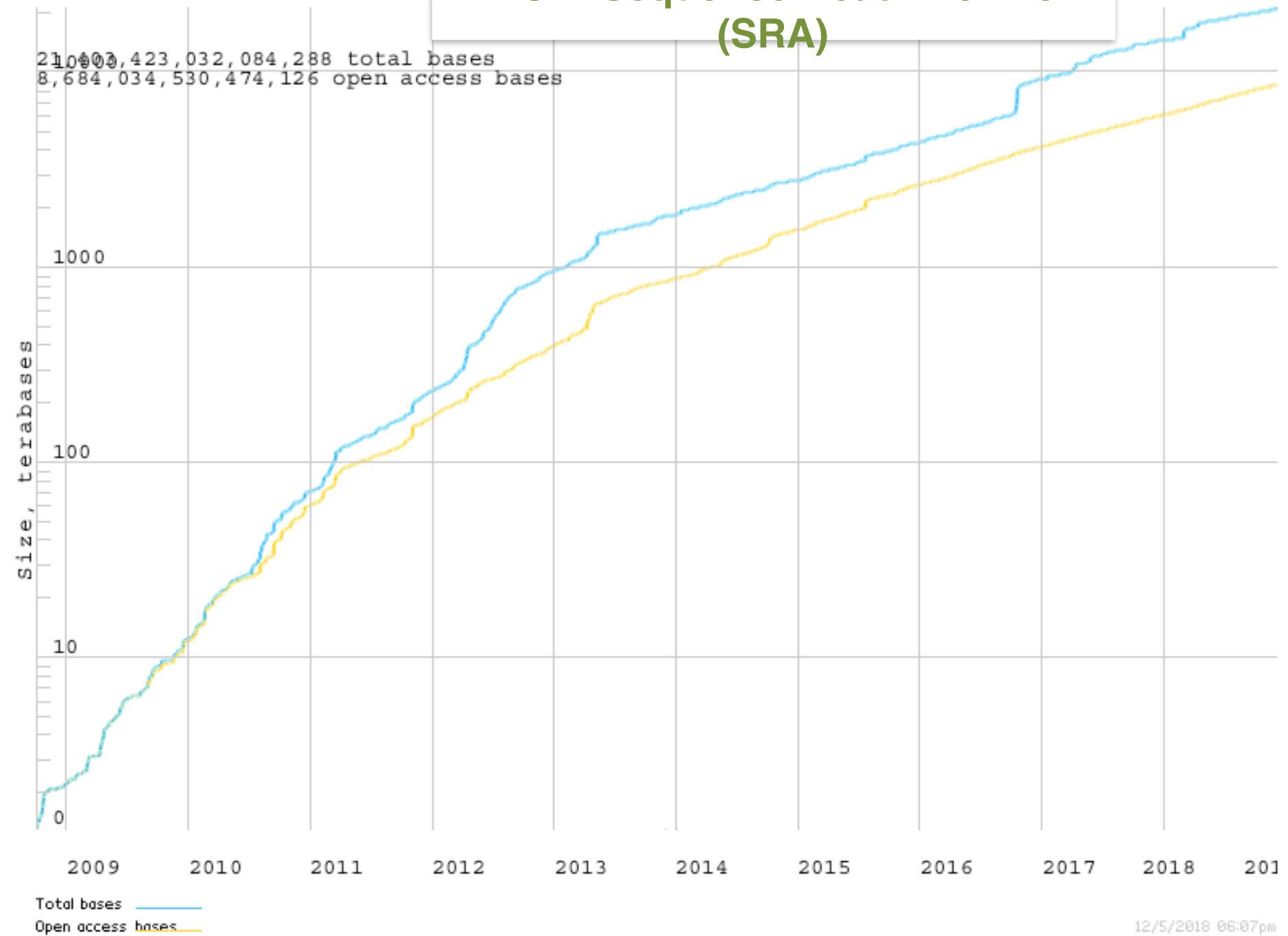
Next Generation Sequencers

High Throughput Sequencing



NCBI Sequence Read Archive (SRA)

21,040,423,032,084,288 total bases
8,684,034,530,474,126 open access bases

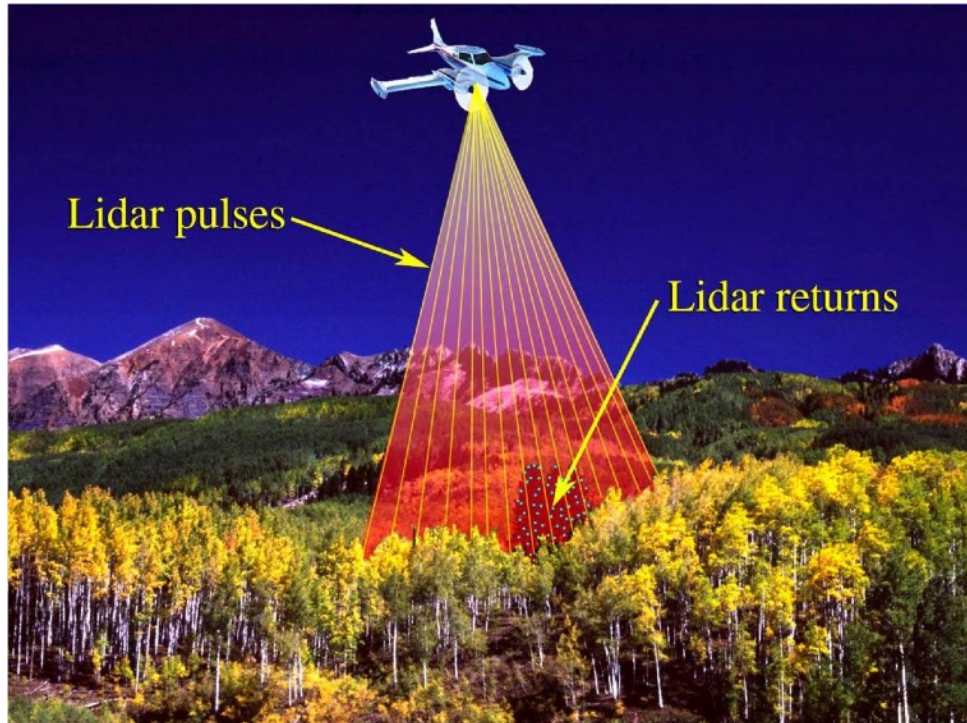


Phenomics



Environmental-omics?

Remote Sensing - LIDAR



Active and passive methods

Monitor the impacts climate change, manage natural resources, and assess research plots

- Shoreline changes
- Ocean temperature
- Soil composition/Sediment transport
- Forest canopy
- Species composition
 - Biodiversity/invasive species

Citizen Science: non-expert contributions

budburst
a project of the Chicago Botanic Garden

My Account

About BudburstGet StartedPlantsProjectsData




Photo courtesy of Jane E. Ogilvie.

Learn How to Observe

Observe Your Plant

Report Your Observation

Download Data

Get Started With Budburst!

Budburst is a national network of citizen scientists monitoring plants as the seasons change. To join, follow these steps: [Learn how to observe](#), [Make an observation](#), and [Report your observation](#).

Learn How to

Watching plants and re of a plant during the gr including leafing, flowe observers benefit from

Citizen Science and Climate Change: Mapping the Range Expansions of Native and Exotic Plants with the Mobile App Leafsnap ^{FREE}

[W John Kress](#), [Carlos Garcia-Robledo](#), [João V B Soares](#), [David Jacobs](#), [Katharine Wilson](#), [Ida C Lopez](#), [Peter N Belhumeur](#)

Challenges of Integratomics: 'Big Data' explained by Data Science

40 ZETTABYTES

[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005



Volume SCALE OF DATA

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**
are shared on Facebook every month



Variety DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be
**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**

**4 BILLION+
HOURS OF VIDEO**
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users



The New York Stock Exchange captures
**1 TB OF TRADE
INFORMATION**
during each trading session



Velocity ANALYSIS OF STREAMING DATA

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure



By 2016, it is projected there will be

**18.9 BILLION
NETWORK
CONNECTIONS**

— almost 2.5 connections per person on earth



**1 IN 3 BUSINESS
LEADERS**

don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Veracity UNCERTAINTY OF DATA

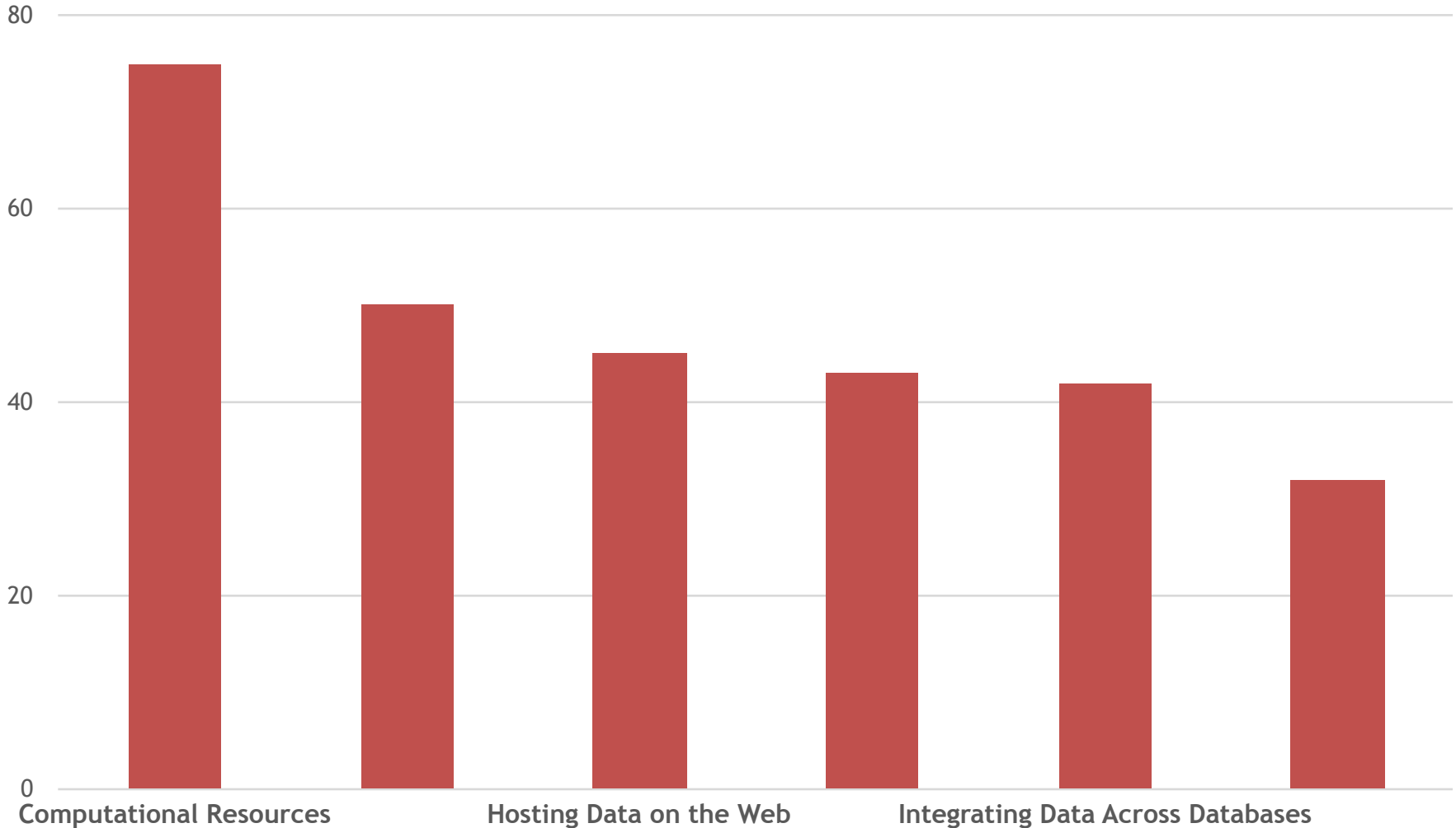
Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



Challenges of Integratomics


Forest Tree Research Community Survey





Improving infrastructure for tree genomic and phenomic data




TreeGenes Database


 **TreeGenes**


TreeGenesCommunityLiteratureSpeciesToolsDownload Data








EXPLORE TREEGENES


**Browse Genomes**
Browse genomes with JBrowse

**CartograTree**
CartograTree is a map-based web app...


**Download Data**
Download data from our FTP site

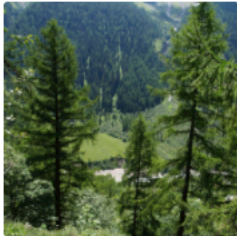
**Sequence Search**
Search sequences with DIAMOND


**Species**
Find your species of interest

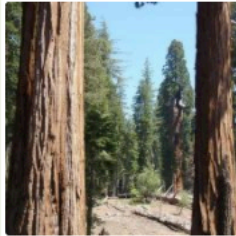
**Submit**
Submit your data to TreeGenes.

FEATURED PROJECTS

**Ash Tree Genomes**

**Silver Fir Genome Project**

**Spruce Genome Project**

**Redwood Genome Project**

MEETINGS

Population, Evolutionary and Quantitative Genetics Conference
May 13 to 16 2018
Madison, USA

Galaxy Community Conference
June 25 to 30 2018
Portland, USA

- View more meetings

LATEST LITERATURE

Stage and Size Structure of Three Species of Oaks In Central Coastal California
(2018) Madroño

TreeGenes Database: Species

treegenesdb.org

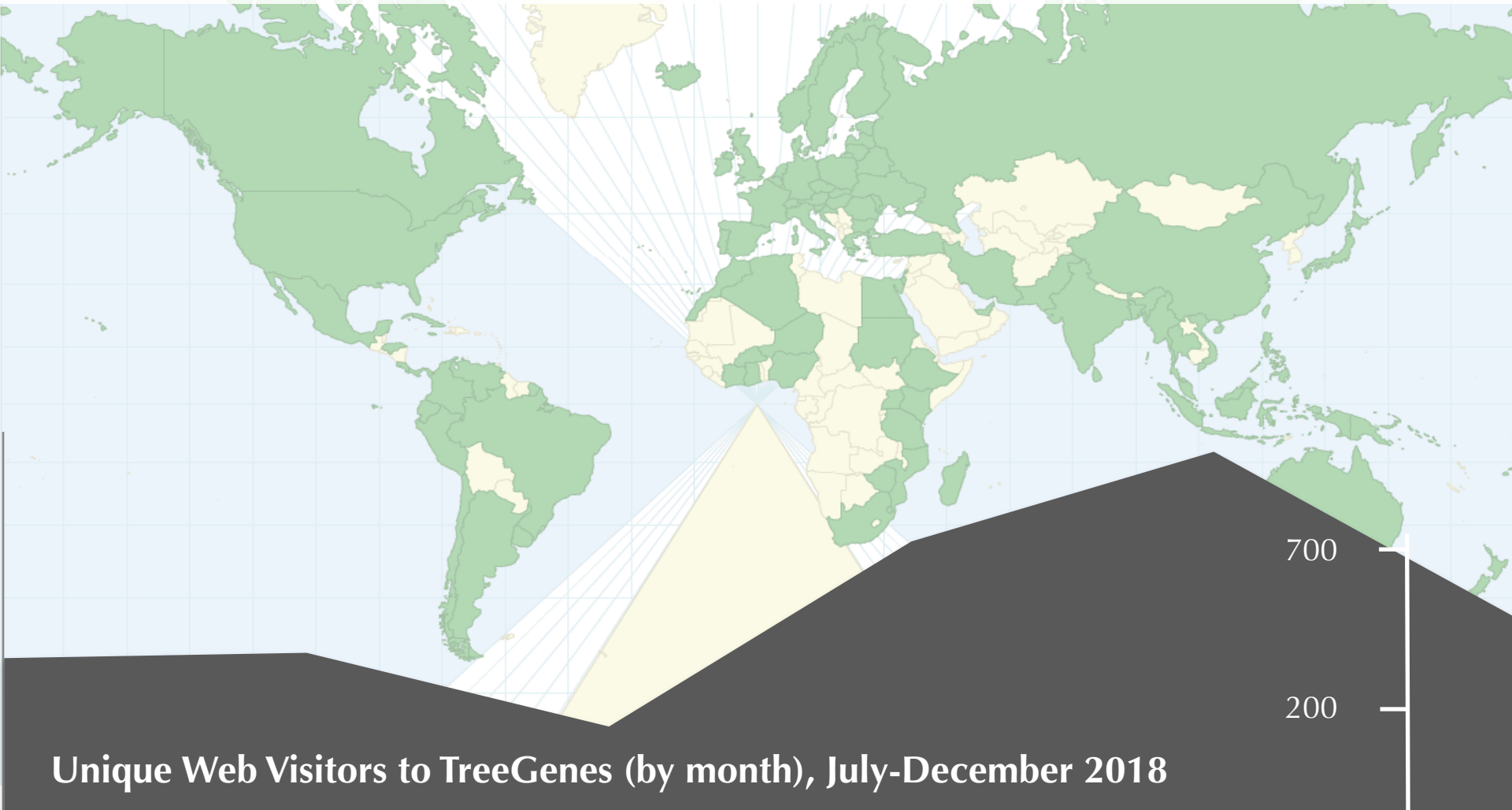


- 1,701 species from 112 genera
 - At least one genetic artifact from each species
- Full genome sequence: 25 species
- Transcriptome/Expression resources: 6,920,817 sequences from 322 species
- 108 genetic maps from 37 species
- **Population studies**
 - Georeferenced trees
 - Extensive genotypic (GBS and array) and phenotypic data

TreeGenes Database: Users

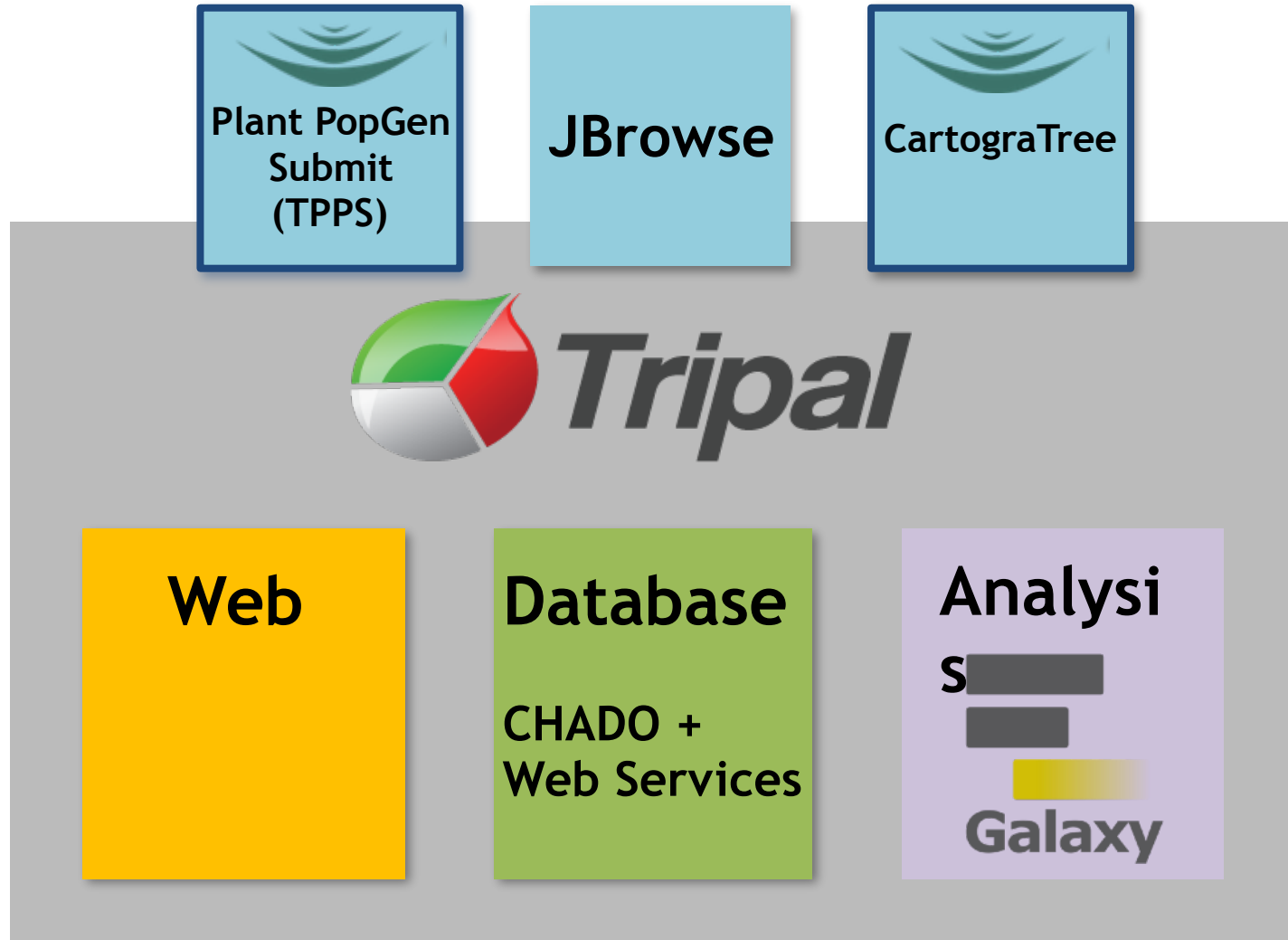
treegenesdb.org

3,100 unique visitors from 116 countries

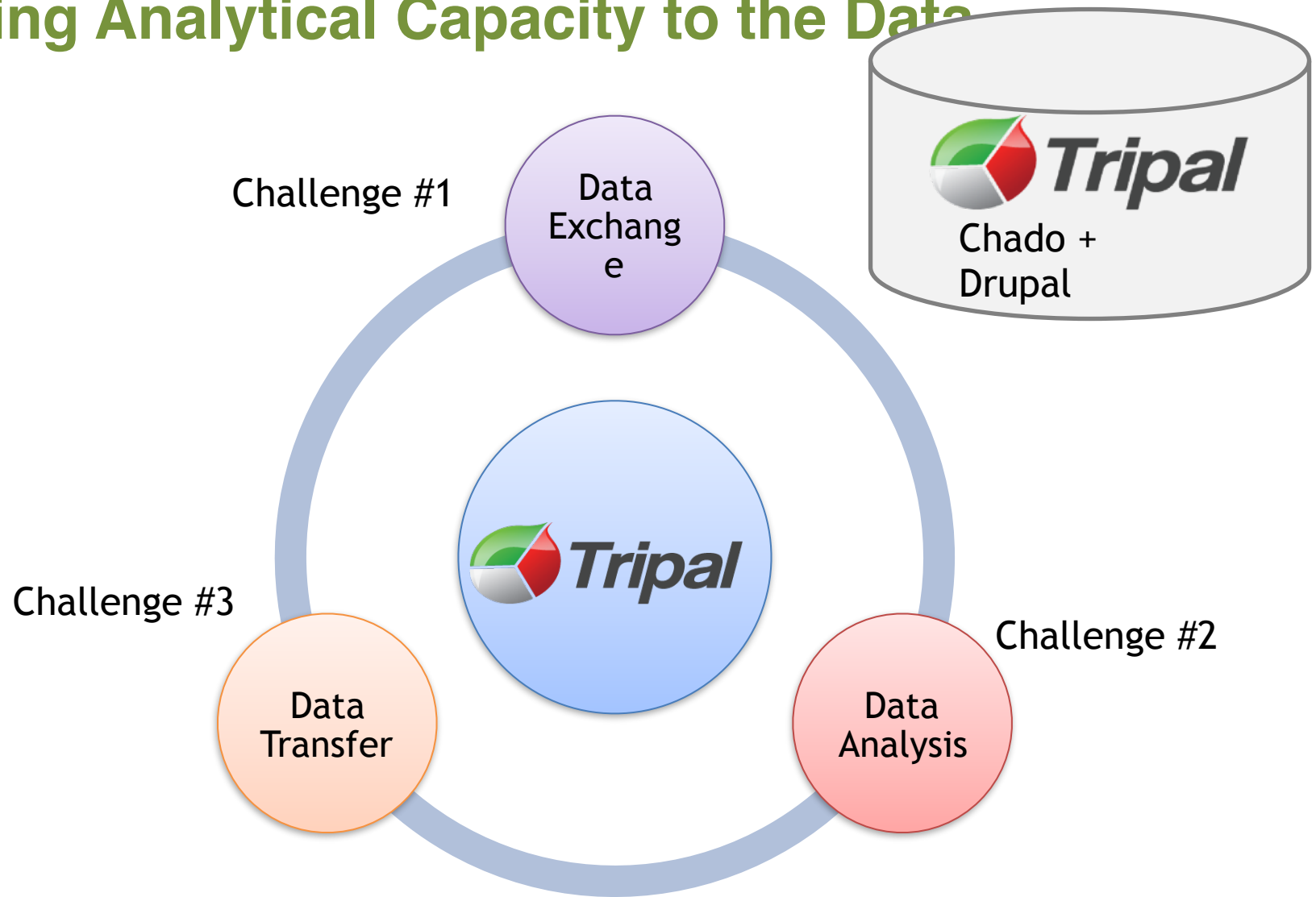


Unique Web Visitors to TreeGenes (by month), July-December 2018

Tripal Framework in TreeGenes



Tripal Gateway Project: Bringing Analytical Capacity to the Data



**TreeGenes hosts georeferenced plants that
can be integrated with environmental metrics**



Scientists require this integration (GxPxE)

- What genotypes contribute to traits related to timber production?
- What genotypes are most adapted to specific elevations/climates for reforestation?
- What genotypes are most resistant to invasive and native pests and pathogens?
- What individuals are best suited for migration within their range in the face of a changing climate?



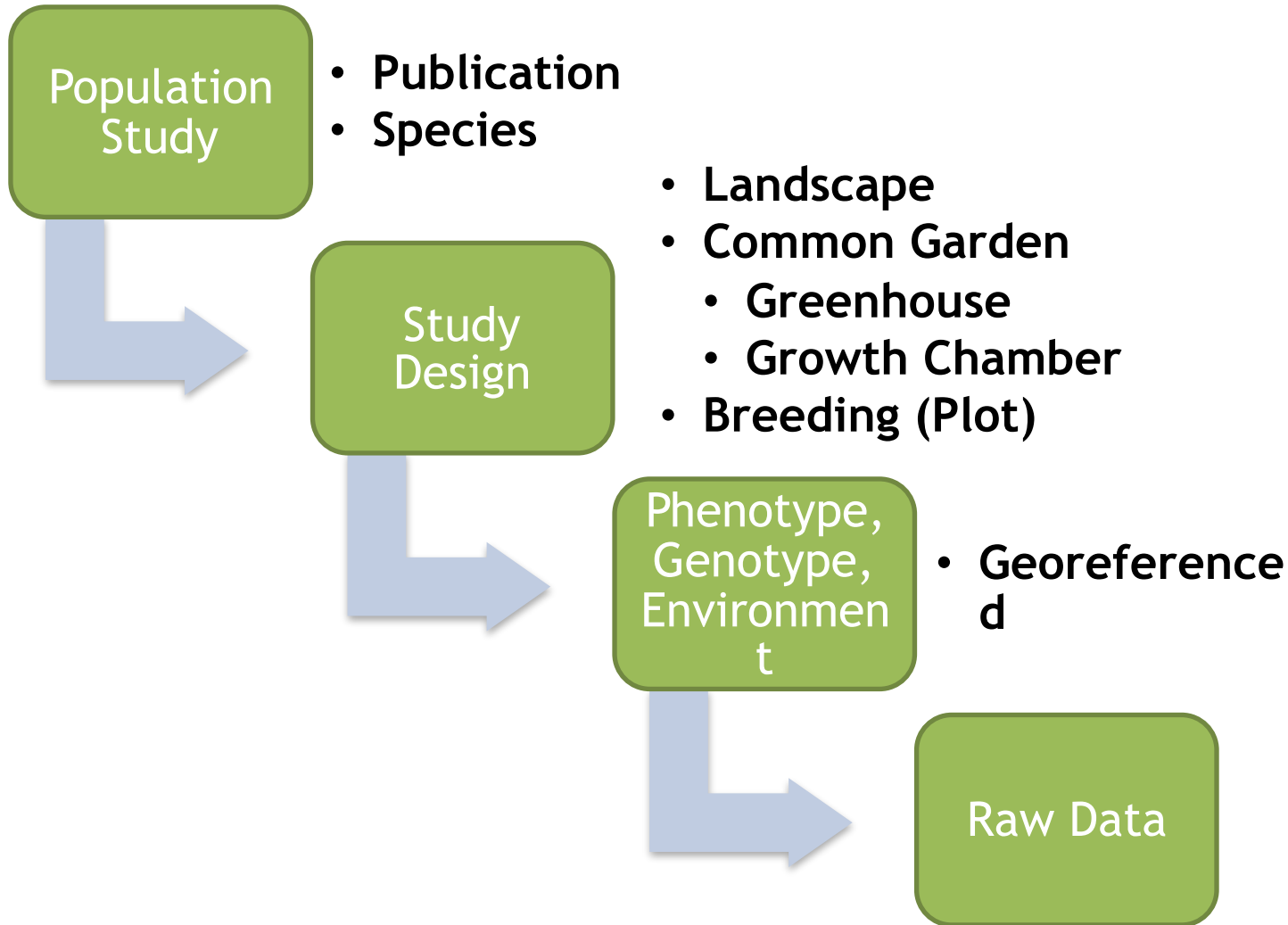
Tripal Plant PopGen Submit (TPPS)

Metadata collection remains sparse and incomplete

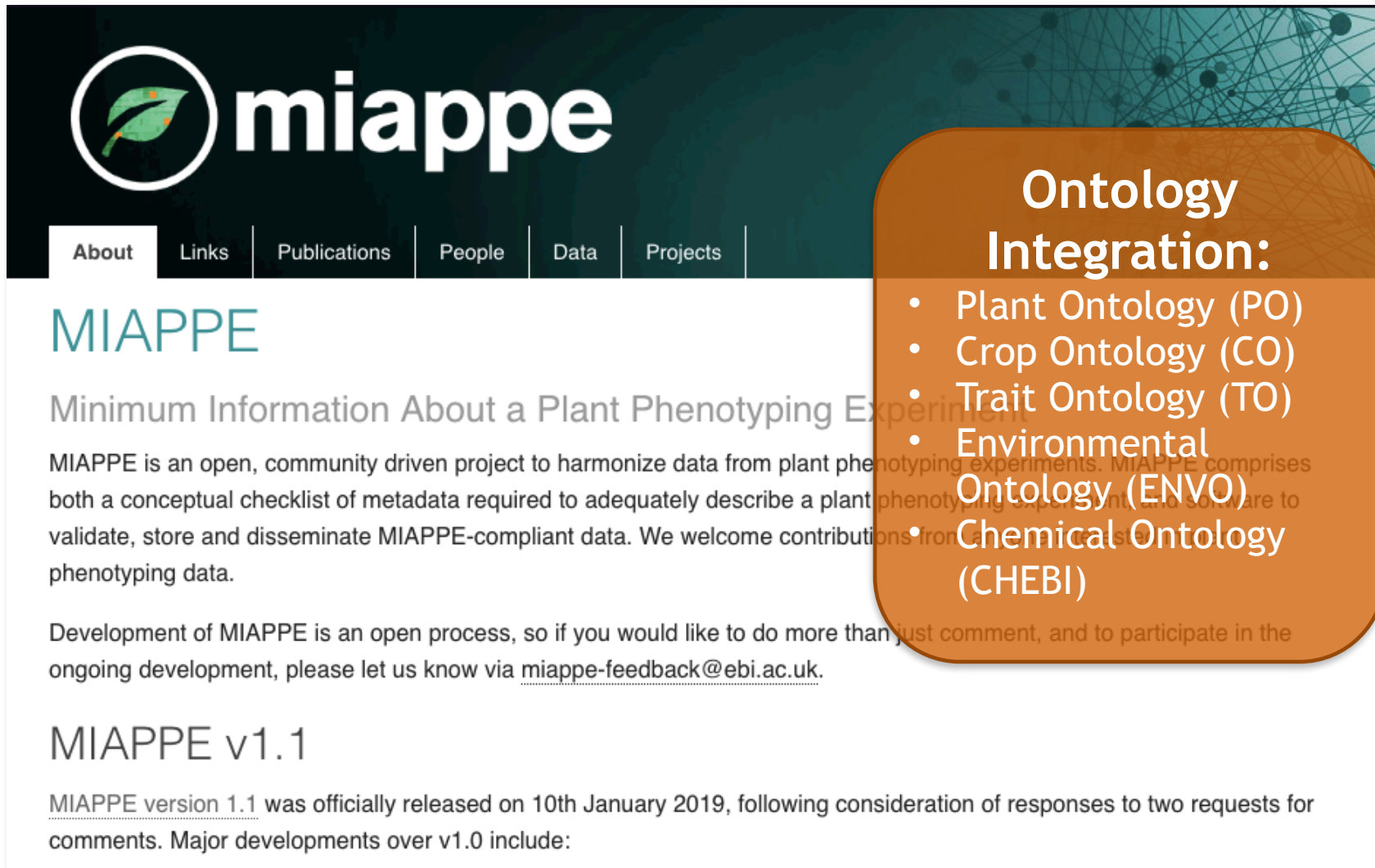
- Long-term accessions (and storage)
- Integration with existing ontological frameworks
- Standards related to data collection
- Integration with primary repositories
- Focus on capturing georeferenced data



Tripal Plant PopGen Submit (TPPS)



Minimal Information About a Plant Phenotyping Experiment (MIAPPE)



The image shows a screenshot of the MIAPPE website. The header features the MIAPPE logo, which consists of a green leaf icon inside a white circle, followed by the text "miappe" in a bold, white, sans-serif font. Below the logo is a navigation bar with links: "About", "Links", "Publications", "People", "Data", and "Projects". The main content area has the heading "MIAPPE" in a large, teal, sans-serif font, followed by the subtitle "Minimum Information About a Plant Phenotyping Experiment". A paragraph describes MIAPPE as an open, community-driven project to harmonize data from plant phenotyping experiments. An orange rounded rectangle is overlaid on the right side of the page, containing the text "Ontology Integration:" followed by a bulleted list of four ontologies: Plant Ontology (PO), Crop Ontology (CO), Trait Ontology (TO), and Environmental Ontology (ENVO). Below this list, the text "Chemical Ontology (CHEBI)" is also visible. At the bottom of the page, the text "MIAPPE v1.1" is displayed, followed by a paragraph stating that MIAPPE version 1.1 was officially released on 10th January 2019, following consideration of responses to two requests for comments. Major developments over v1.0 include:

miappe

About | Links | Publications | People | Data | Projects

MIAPPE

Minimum Information About a Plant Phenotyping Experiment

MIAPPE is an open, community driven project to harmonize data from plant phenotyping experiments. MIAPPE comprises both a conceptual checklist of metadata required to adequately describe a plant phenotyping experiment, and software to validate, store and disseminate MIAPPE-compliant data. We welcome contributions from the plant phenotyping community.

Development of MIAPPE is an open process, so if you would like to do more than just comment, and to participate in the ongoing development, please let us know via miappe-feedback@ebi.ac.uk.

MIAPPE v1.1

MIAPPE version 1.1 was officially released on 10th January 2019, following consideration of responses to two requests for comments. Major developments over v1.0 include:

- Plant Ontology (PO)
- Crop Ontology (CO)
- Trait Ontology (TO)
- Environmental Ontology (ENVO)
- Chemical Ontology (CHEBI)

Tripal Plant PopGen Submit (TPPS)

Phenotype Metadata File: Please upload a file containing columns with the name, attribute, description, and units of each of your phenotypes: *

 phenotype metadata.xlsx **REMOVE**

File Upload empty field: NA

By default, TPPS will treat cells with the value "NA" as empty. If you used a different empty value indicator, please provide it here.

▼ DEFINE DATA

Please define which columns hold the required data: Phenotype name

name	attribute	units	
Phenotype Name/Identifier	Attribute	Units	Description
phenotype 1	age	years	quantitative
phenotype 2	age	years	quantitative
phenotype 3	age	years	quantitative

**TreeGenes Data
Repository Accession
(TGDR####) -> DOI**

Phenotype file below:

Phenotype 2 Attribute: *

Some examples of attributes include: "amount", "width", "mass density", "area", "height", "age", "broken", "time", "color", "composition", etc.

Phenotype 2 Description: *

Please provide a short description of Phenotype 2

Phenotype 2 Units: *

☐ Humidity regime

[Click here to view trees on map!](#)



Map data ©2018 Google, IN Terms of Use

FAIR
Findable
Accessible



TreeGenes to CartograTree

TPPS/TGDR DETAILS FOR TGDR001

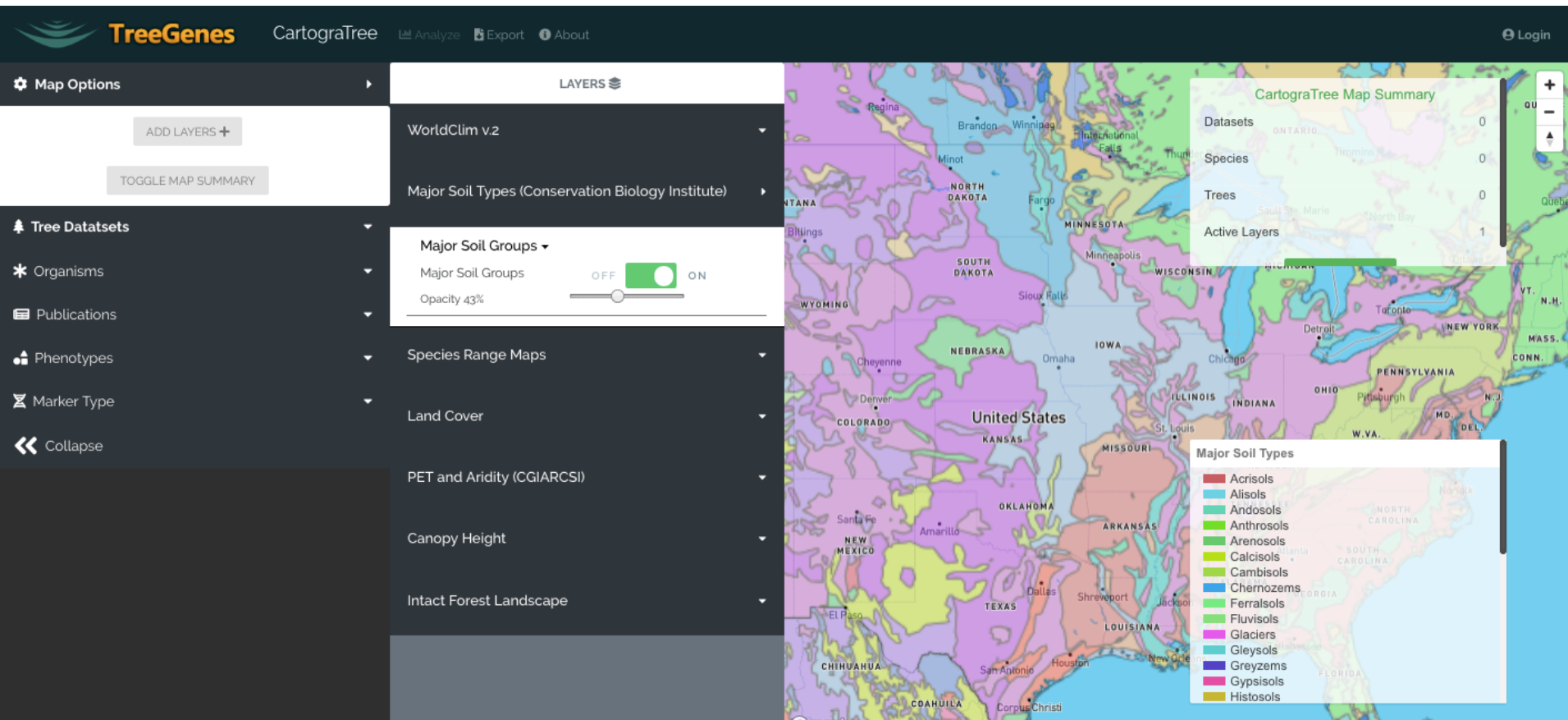
Attribute	Details
Accession	TGDR001
Title	Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (<i>Populus trichocarpa</i> , Salicaceae) secondary xylem.
Species	<i>Populus trichocarpa</i>
Study Type	GxP
File Downloads	<ul style="list-style-type: none">✓ ASSOCIATION RESULTS FILE 1 FILE✓ GENOTYPES SNP✓ GPS COORDINATES✓ HAPLOTYPE DATA FILE✓ PHENOTYPES✓ PHENOTYPES DEFINITIONS
CartograTree	View in CartograTree
Tree Count	1376
Phenotype Count	4032
Unique Phenotypes	3



The screenshot displays the TreeGenes CartograTree web application interface, which is used for mapping and analyzing tree distribution data. The interface is organized into several key components:

- Left Sidebar (Navigation & Settings):**
 - Map Options:** Includes a 'Map Summary' toggle (currently OFF) and a 'Tree Dataset Sources' section with toggles for TreeGenes, TreeSNAP, and DRYAD (all currently ON).
 - Organisms:** A section for selecting taxonomic levels, with 'Family' currently selected. Below this are filters for 'Family', 'Genus', and 'Species'.
 - Publications:** A section for viewing publication-related data.
- Central Layers Panel:** A panel titled 'LAYERS' that allows users to manage the data layers displayed on the map. The layers listed include:
 - WorldClim v2:** Sub-layers for Precipitation, Temperature, Solar Radiation (with a sub-panel for January solar radiation and a 39% opacity slider), Wind Speed, and Water Vapor.
 - Major Soil Types (Conservation Biology Institute)**
 - Species Range Maps**
 - Land Cover**
 - PET and Aridity (CGIARCSI)**
 - Canopy Height**
- Main Map Area:** A map of the Southeastern United States showing the distribution of tree species. A pop-up window for *Celtis occidentalis* (ID: treesnap.194) is displayed, showing its location, coordinates (Latitude: 36.959, Longitude: -86.338), and a 'SAVE' button. The map also shows a 'Map Summary' panel on the right with statistics for Publication Datasets (0), Species (195), Trees (34220), and Active Layers (2).

CartograTree: Integrating environmental layers with georeferenced trees



CartograTree: Integrating environmental layers with georeferenced trees

The screenshot displays the CartograTree web application interface. The top navigation bar includes the TreeGenes logo, the CartograTree title, and links for 'Analyze' and 'About'. The left sidebar contains several interactive panels:

- Phenotypes**: A panel with a 'Phenotype' section featuring an 'Additive' toggle switch (currently OFF) and a search input field.
- Phenotype & Trait Ontology**: A panel with a search input field.
- Plant Ontology**: A panel with a search input field and a list of selected ontologies: 'inflorescence bud' (306), 'leaf' (1201), and 'bud burst stage' (435).
- Marker Type**: A panel with a 'Collapse' button.

The main map area shows a geographical view of Europe, with tree markers placed at various locations. A detailed information popup is displayed for a tree marker in the Netherlands, showing the following data:

- Coordinates**: Latitude: 49.25 | Longitude: 3.1, Coordinate type: exact
- Species**: *Fagus sylvatica* (angiosperm)
- Family**: Fagaceae
- ID**: TGDR069-190141
- Actions**: DEL (button), ADD (button)
- Environmental Data**: 1 of 15 (indicated by arrows and the TreeGenes logo)

CartograTree: Save searches locally and select for meta-analysis

The image displays the CartograTree web application interface, which is a Forest Tree Map Utility. The interface is divided into several sections:

- Left Panel (Phenotypes):** Contains a sidebar with a 'Phenotypes' section. It includes a 'Phenotype' toggle (OFF/ON), an 'Additive' button, and a 'Phenotype & Trait Ontology' section with a 'Plant Ontology' section. The 'Plant Ontology' section lists 'Inflorescence bud' (305), 'leaf' (1201), and 'bud burst stage' (435).
- Map:** A map of France showing the location of the study area. The map is labeled 'France' and shows major cities like Paris, Orleans, and Troyes.
- CartograTree Analysis Modal:** A central modal window titled 'CartograTree Analysis' with the subtitle 'A Forest Tree Map Utility'. It contains the following sections:
 - Select State of Map:** A button to select the state of the map.
 - Additional Options:** A button to access additional options.
 - Summary and Confirm:** A button to confirm the analysis.
 - Current Map State:** A table showing the current map state.

# Trees	# Species	# Layers
1201	0	1
 - Load Saved Searches:** A table showing saved searches.

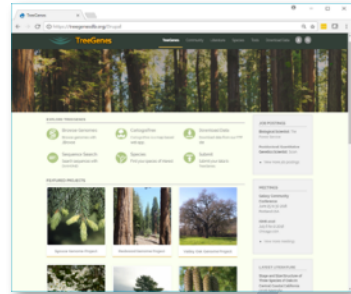
Date	Title	# Trees	# Species
 - Analysis type:** A dropdown menu set to 'Landscape GxE'.
 - Publications Selected:** A table showing selected publications.

ID	Title	Author	Year	# Trees	Study type	Status
 - Environmental variables:** A table showing environmental variables.

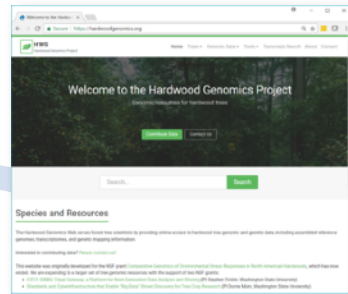
Layer name	Source	Environmental values
Solar radiation January	http://worldclim.org/version2	<input checked="" type="checkbox"/> Solar_radiation
 - Buttons:** 'BACK', 'NEXT', and 'CLOSE' buttons.



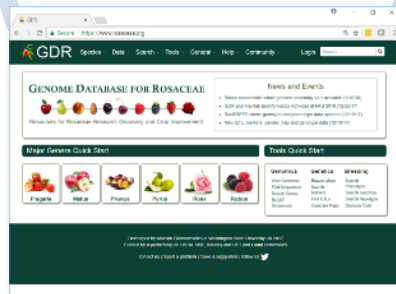
Bringing Analytical Capacity to the Data



Tree Genes

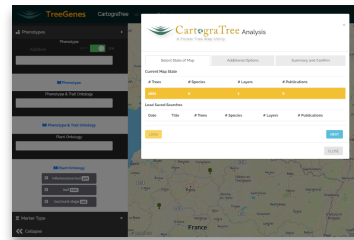


Hardwood Genomics



Genome Database for Rosaceae

CartograTree



Galaxy
PROJECT



Galaxy: Open Source Web-based Platform for Bioinformatic Analysis

The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo, a search bar, and links for Analyze Data, Workflow, Visualize, Shared Data, Help, Login or Register, and a user status indicator (Using 0%).

Tools Panel (Left): A sidebar with a search bar and a list of tool categories: 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, and NGS: Mothur.

Main Content Area:

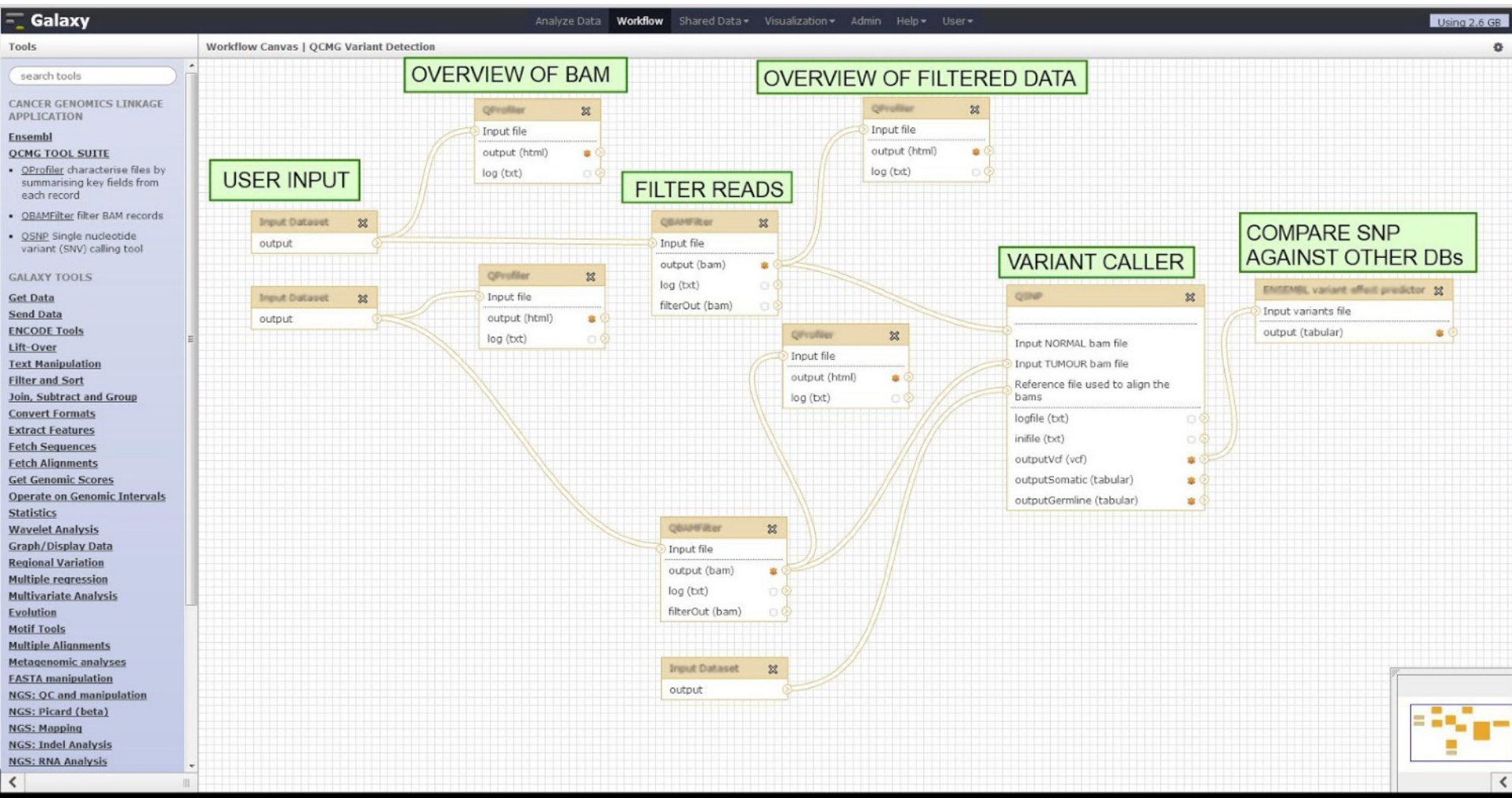
- Text:** Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our help resources. You can install your own Galaxy by following the [tutorial](#) and choose from thousands of tools from the [Tool Shed](#).
- Galaxy 101:** An introduction to Galaxy tutorial, featuring a diagram of the Galaxy Training Network.
- Tweets:** A section titled "Tweets by @galaxyproject" showing a tweet from the Galaxy Project (@galaxyproject) about Galaxy Platform News, including links to @EGI_einfra, ChemFlow, ChIP-Seq Docker, @usegalaxy, and @GalaxyAustralia.
- Galaxy Platforms News:** A section titled "Galaxy Platforms News" with a link to the Galaxy Platform News page.

History Panel (Right): A sidebar with a search bar and a section titled "Unnamed history (empty)". A message states: "This history is empty. You can [load your own data](#) or [get data from an external source](#)".

Footer: Logos for PennState, Johns Hopkins University, Oregon Health & Science University, TACC, and CYVERSE.



Galaxy: Open Source Web-based Platform for Bioinformatic Analysis



Workflows for Landscape Genomics: Integrating across diverse datasets

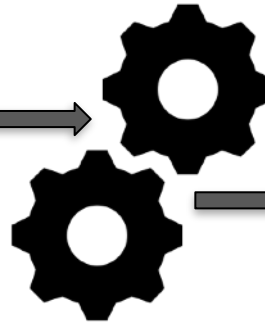
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SN
TREE1	0	0	0	0	0	0	0	0	0	0
TREE2	1	0	0	0	0	0	0	0	0	0
TREE3	1	0	0	0	0	0	0	1	1	0
TREE4	1	0	0	0	0	0	0	0	0	0
TREE5	1	0	0	0	0	0	0	0	0	0
TREE6	1	0	0	0	0	0	0	1	0	0
TREE7	0	0	0	0	0	0	0	1	1	0
TREE8	0	0	0	0	0	0	0	0	1	0
TREE9	1	0	1	1	0	1	0	0	1	0
TREE10	0	0	0	1	0	0	0	0	0	0

Genotypic data

	PC1	PC2	PC3	PC4	PopStr
TREE1	-8.648228102	3.173037266	3.66273579	-0.177338137	-8.554180345
TREE2	-8.554180345	2.43752863	0.93398487	0.04374811	0.37830172
TREE3	3.94101741	0.4768156	0.093398487	0.04374811	0.37830172
TREE4	0.4768156	0.093398487	0.04374811	0.37830172	0.37830172
TREE5	0.093398487	0.04374811	0.37830172	0.37830172	0.37830172
TREE6	0.04374811	0.37830172	0.37830172	0.37830172	0.37830172
TREE7	0.37830172	0.37830172	0.37830172	0.37830172	0.37830172
TREE8	0.37830172	0.37830172	0.37830172	0.37830172	0.37830172
TREE9	-8.570749832	2.761219441	4.189358619	0.120877968	-9.732114439
TREE10	-8.570749832	2.761219441	4.189358619	0.120877968	-9.732114439

Environmental data

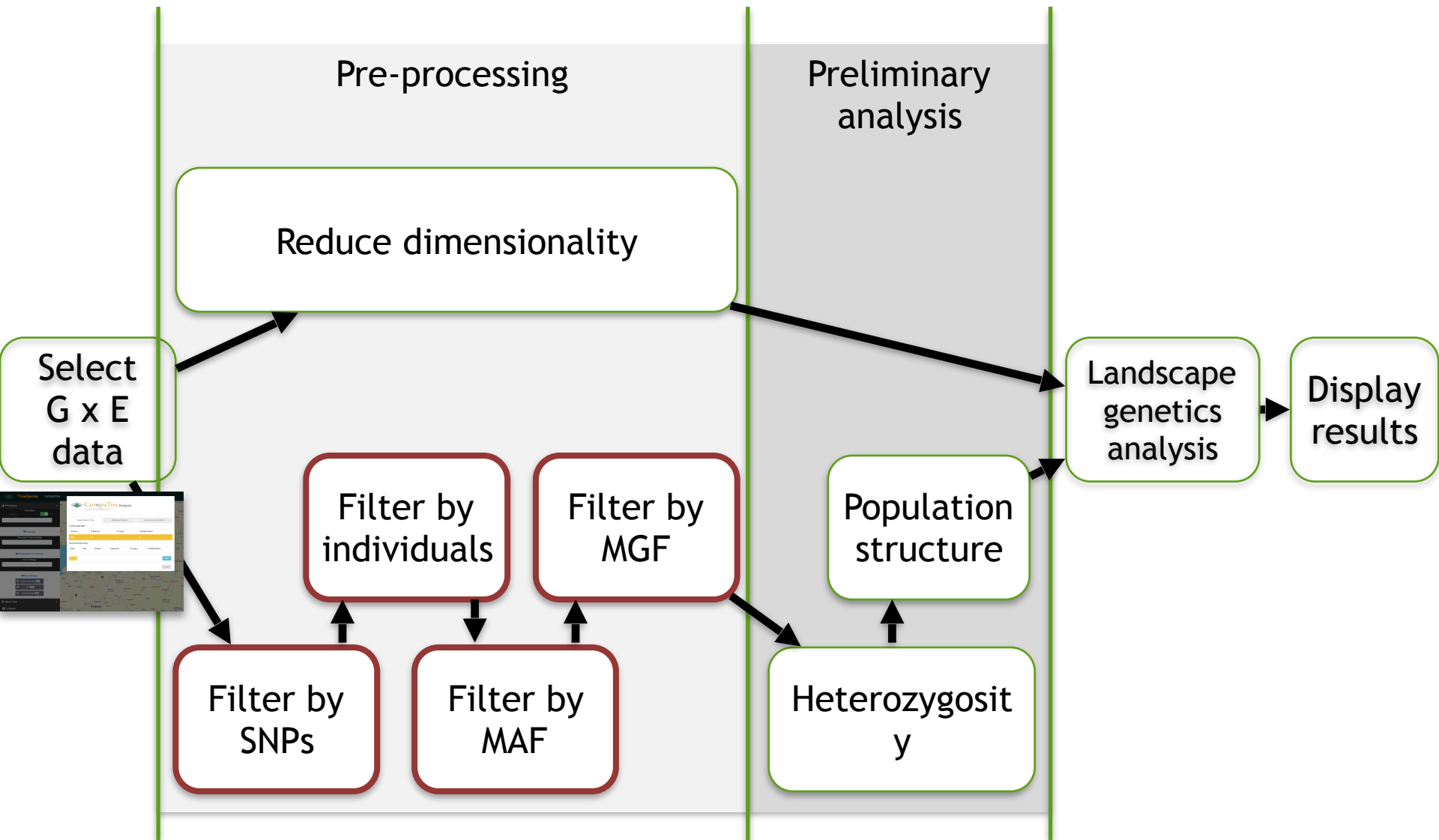
Galaxy



Marker	Env_1	Env_2	Loglikelihood	Gscore	WaldScore	NumError	Efron	McFadden	M
SNP1	PC1	PC3	-439.2146473	31.677384	30.54585431	0	0.192739468	0.034806126	0
SNP2	PC1	PopStr	-351.0	0	0	0	0.248893702	0.037839423	0
SNP3	PC4	PopStr	-350.0	0	0	0	0.231231003	0.041423415	0
SNP4	PC3	PopStr	-376.0	0	0	0	0.345328702	0.029909067	0
SNP5	PC1	PC3	-495.0	0	0	0	0.104945853	0.020848599	0
SNP6	PC1	PC4	-439.0	0	0	0	0.093285957	0.020636018	0
SNP7	PC3	PopStr	-423.0	0	0	0	0.267726345	0.023629761	0
SNP8	PC3	PopStr	-355.0	0	0	0	0.20534412	0.025734199	0
SNP9	PC1	PC3	-454.1183637	18.47638759	17.29764504	0	0.045307766	0.019937548	0

Correlated markers

Workflows: Executed in Galaxy with metadata



Future Development:

CartograTree -> CartograPlant

- Developing models to load balance analysis (Galaxy/TACC)
- Leverage Cyverse DataStore to share and store data
- Develop more workflows (association genetics)
- Robust platform for additional plant species



Plant Computational Genomics Lab, University of Connecticut

- Nic Herndon
- Emily Grau
- Sean Buehler
- Ronald Santos
- Risharde Ramnath
- Peter Richter

Washington State University

- Stephen Ficklin
- Doreen Main

University of Tennessee

- Margaret Staton
- Ming Chen

