CartograPlant: Cyberinfrastructure to improve plant health and productivity in the context of a changing climate

Irene Cobo Simón, PhD.
Postdoctoral Fellow
Institute of Forest Science (ICIFOR-INIA-CSIC, Spain)
What is CartograPlant?

https://cartograplant.org/

BACKGROUND

INTEGRATES

VISUALIZES

ANALYZES

Genotypes

Phenotypes

Environments

From georeferenced plants (including forest trees)
Why is CartograPlant relevant and timely?

- Climate change is threatening plant health and productivity

Green ashes affected by the pest emerald ash borer
Why is CartograPlant relevant and timely?

- **Climate change** is threatening **plant health and productivity**
  - Can **plant breeding** keep pace with the **rate and direction of environmental change**?

**BACKGROUND**

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

Green ashes affected by the pest emerald ash borer
Why is CartograPlant relevant and timely?

- **Climate change** is threatening **plant health and productivity**
  - Can **plant breeding** keep pace with the **rate and direction** of environmental change?
  - Increasing **invasive pests and pathogens**
Why is CartograPlant relevant and timely?

- **Climate change** is threatening **plant health and productivity**
  - Can **plant breeding** keep pace with the **rate and direction of environmental change**?
  - Increasing **invasive pests and pathogens**
- **Illegal logging and deforestation** (forest trees)
Why is CartograPlant relevant and timely?

- Match between genotypes, phenotypes and new environments

Background: Green ashes affected by the pest emerald ash borer.
Why is CartograPlant relevant and timely?

- Match between genotypes, phenotypes and new environments
  - Candidate gene identification (resilience)
Why is CartograPlant relevant and timely?

- Match between genotypes, phenotypes and new environments
  - Candidate gene identification (resilience)
  - Timber origin identification, using chemical, genetic and anatomic tree data

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

Green ashes affected by the pest emerald ash borer
Why is CartograPlant relevant and timely?

- Match between genotypes, phenotypes and new environments
  - Candidate gene identification (resilience)
  - Timber origin identification, using chemical, genetic and anatomic tree data

Tools that collect, integrate and facilitate these data, such as CartograPlant, are critical.
DATA INTEGRATION

DATA TYPES INTEGRATED IN CARTOGRAPPLANT

1. RAW DATA

- GENOTYPIC
- PHENOTYPIC
- ENVIRONMENTAL

BACKGROUND | DATA INTEGRATION | DATA VISUALIZATION | DATA ANALYSIS
DATA TYPES INTEGRATED IN CARTOGRAPLANT

1. RAW DATA

- GENOTYPIC
- PHENOTYPIC
- ENVIRONMENTAL

How can these disparate data types, from different studies, be integrated?
DATA INTEGRATION

DATA TYPES INTEGRATED IN CARTOGRAPLANT

1. RAW DATA
- GENOTYPIC
- PHENOTYPIC
- ENVIRONMENTAL

2. METADATA + ONTOLOGIES + STANDARDS

BACKGROUND
DATA INTEGRATION
DATA VISUALIZATION
DATA ANALYSIS
DATA INTEGRATION

DATA TYPES INTEGRATED IN CARTOGRAPLANT

1. RAW DATA
   - GENOTYPIC
   - PHENOTYPIC
   - ENVIRONMENTAL

2. METADATA + ONTOLOGIES + STANDARDS

Direct submission of studies
Tripal Plant PopGen Submit (TPPS) pipeline
DATA INTEGRATION

DATA TYPES INTEGRATED IN CARTOGRAPPLANT

1. RAW DATA
   - GENOTYPIC
   - PHENOTYPIC
   - ENVIRONMENTAL

2. METADATA + ONTOLOGIES + STANDARDS

Direct submission of studies

Tripal Plant PopGen Submit (TPPS) pipeline

Biocuration efforts of affiliated databases

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS
Direct submission: Tripal Plant PopGen Submit (TPPS) pipeline

- Genotype, phenotype, environmental data and metadata and provides a DOI

https://treegenesdb.org/tpps
Direct submission: Tripal Plant PopGen Submit (TPPS) pipeline

- Genotype, phenotype, environmental data and metadata and provides a DOI
- Population genomics, association mapping, and landscape genomic studies

https://treegenesdb.org/tpps
DATA INTEGRATION

Direct submission: Tripal Plant PopGen Submit (TPPS) pipeline

• Genotype, phenotype, environmental data and metadata and provides a DOI

• Population genomics, association mapping, and landscape genomic studies

• Metadata is collected using ontologies and standards (MIAPPE)

https://treegenesdb.org/tpps
Direct submission: Tripal Plant PopGen Submit (TPPS) pipeline

- Genotype, phenotype, environmental data and metadata and provides a DOI
- Population genomics, association mapping, and landscape genomic studies
- Metadata is collected using ontologies and standards (MIAPPE)
- Ensures the FAIRness of the data

https://treegenesdb.org/tpps
Direct submission: Tripal Plant PopGen Submit (TPPS) pipeline

- Genotype, phenotype, environmental data and metadata and provides a DOI
- Population genomics, association mapping, and landscape genomic studies
- Metadata is collected using ontologies and standards (MIAPPE)
- Ensures the FAIRness of the data
- Supports standard genotyping file formats (VCF)

https://treegenesdb.org/tpps
The vast majority of the TPPS submitted studies (and their associated genotype, phenotype and environmental data) are available in CartograPlant thanks to our Biocuration team.
Environmental layers

950 environmental layers are now available in CartograPlant

<table>
<thead>
<tr>
<th>Data Integration</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Data</td>
<td>National Forests</td>
</tr>
<tr>
<td>Ecoregions</td>
<td>Low Impact Areas.</td>
</tr>
<tr>
<td>Soil Type</td>
<td>Population Density</td>
</tr>
<tr>
<td>Forest Fragmentation</td>
<td>Intact Forests</td>
</tr>
<tr>
<td>Neon Field Stations</td>
<td>World Forest ID Data</td>
</tr>
<tr>
<td>Seed Zones</td>
<td>Protected Areas</td>
</tr>
<tr>
<td>Biodiversity Hotspots</td>
<td>Human Impact</td>
</tr>
<tr>
<td>Pet/Aridity</td>
<td>Biotic Damage</td>
</tr>
<tr>
<td>Canopy Height</td>
<td>NDVI (Plant Health)</td>
</tr>
<tr>
<td>Species Ranges</td>
<td>Forest Fragmentation</td>
</tr>
<tr>
<td>Land Cover</td>
<td>World Forest ID Data</td>
</tr>
<tr>
<td>Species Ranges</td>
<td>Population Density</td>
</tr>
<tr>
<td>Species Ranges</td>
<td>Population Density</td>
</tr>
</tbody>
</table>

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS
CartograPlant current statistics

- Plants 8,439,968
- Species 635
- Genera 277
- Countries 43
- Studies 313
- Genotypes 771,763,817
- Phenotypes 1,741,822
- Environmental layers 950
DATA VISUALIZATION

LEFT PANEL
Action panel to interact with the map and plants, located to the left of the screen.
DATA VISUALIZATION

LEFT PANEL
Action panel to interact with the map and plants, located to the left of the screen.

RIGHT PANEL
An interactive map, showing the selected plants and environmental layers on the left panel.

BACKGROUND
DATA INTEGRATION
DATA VISUALIZATION
DATA ANALYSIS
DATA VISUALIZATION

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

https://cartograplant.org/
DATA VISUALIZATION

Study Associated

A range-wide herbarium-derived dataset indicates high levels of gene flow in black cherry (Prunus serotina)

Konrade, Lauren 2019  View Additional Details

STUDY FILE DOWNLOADS

Phenotype Accession
Genotype SSRs/cpSSRs

Genotypic Data

<table>
<thead>
<tr>
<th>Marker Name</th>
<th>Genotype</th>
<th>Marker Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>509_48</td>
<td>NA</td>
<td>microsatellite</td>
</tr>
<tr>
<td>506_34</td>
<td>230</td>
<td>microsatellite</td>
</tr>
<tr>
<td>510_50</td>
<td>175</td>
<td>microsatellite</td>
</tr>
</tbody>
</table>

https://cartograplant.org/
DATA ANALYSIS

DATA INTEGRATION

DATA VISUALIZATION

BACKGROUND

https://cartograpplant.org/
DATA ANALYSIS

BACKGROUND
DATA INTEGRATION
DATA VISUALIZATION
DATA ANALYSIS

https://cartograplant.org/
DATA ANALYSIS

- Background
- Data Integration
- Data Visualization
- Data Analysis

https://cartograplant.org/

- plant height
- ratio of average warmest month temperature to average summer precipitation
- volume
- whole plant mass
- wood carbon 13 content

- 3672 phenotypes: Overlaps with all studies
- 918 phenotypes: Overlaps with all studies
- 2754 phenotypes: Overlaps with all studies
- 918 phenotypes: Overlaps with all studies
- 918 phenotypes: Overlaps with all studies

- bud break
- bud set
- carbon to nitrogen ratio

- Adjust thresholds: frost free days
- Adjust thresholds: leaf chlorophyll content
- Adjust thresholds: leaf nitrogen 15 content

- Save adjustment
DATA ANALYSIS

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

https://cartograplant.org/
DATA ANALYSIS

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

https://cartograplant.org/
DATA ANALYSIS

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

https://cartograplant.org/
DATA ANALYSIS

https://cartograplant.org/
Histograms to help decide the quality filtering thresholds available soon!

https://cartograplant.org/
DATA ANALYSIS

DATA INTEGRATION

DATA VISUALIZATION

BACKGROUND

https://cartograplant.org/
Population structure calculation (PCA and DAPC) and visualization (fastSTRUCTURE, PCA and DAPC) available soon!!!
DATA ANALYSIS

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

https://cartograplant.org/
DATA ANALYSIS

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

https://cartograplant.org/
DATA ANALYSIS

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION

DATA ANALYSIS

GWAS with EMMAX
- Landscape genomics with Sambada
- Landscape genomics with Bayenv
- Multiple testing correction (FDR)
- BLUP calculation (phenotypes)
- Meta-analysis with METASOFT
- Meta-analysis with PLINK

Other analytic workflows available soon!

https://cartograpplant.org/
DATA ANALYSIS

BACKGROUND

DATA INTEGRATION

DATA VISUALIZATION
CARTOGRAPHY WORKFLOWS OVERVIEW

1. SNP QUALITY FILTERING
   2. SNP IMPUTATION
   3. SNP REMAPPING

4. POPULATION STRUCTURE
   BayeScan, LD, fastSTRUCTURE

5. GWAS
   EMMAX

5.1. BLUPs

5.2. FDR

6. LANDSCAPE
   GENOMICS
   Bayenv, Samβada

6.1. Multicollinearity

7. META-ANALYSIS

METADATA + ONTOLOGIES + STANDARDS

0. MEGA-ANALYSIS

SUMMARY STATISTICS

CARTOGRAPHY PLANT WORKFLOWS OVERVIEW

1. SNP QUALITY FILTERING
2. SNP IMPUTATION
3. SNP REMAPPING

4. POPULATION STRUCTURE
   BayeScan, LD, fastSTRUCTURE

5. GWAS
   EMMAX

5.1. BLUPs

5.2. FDR

6. LANDSCAPE
   GENOMICS
   Bayenv, Samβada

6.1. Multicollinearity

PHENOTYPE

ENVIRONMENT

GENOTYPE

0. MEGA-ANALYSIS

SUMMARY

METADATA +

ONTOGOLOGIES +

STANDARDS


CONCLUSIONS

- **CartograPlant** is intended to serve as a community resource for Plant Molecular Ecology.
• **CartograPlant** is intended to serve as a community resource for Plant Molecular Ecology.

• These flexible analytic workflows allow to analyze a diversity of data types (e.g. SNPs, SSR) and experimental designs (e.g. natural populations, common garden) and facilitate a **diversity of biological questions in CartograPlant**.
• **CartograPlant** is intended to serve as a community resource for Plant Molecular Ecology.

• These flexible analytic workflows allow to analyze a diversity of data types (e.g. SNPs, SSR) and experimental designs (e.g. natural populations, common garden) and facilitate a diversity of biological questions in CartograPlant.

• Mega and meta-analysis take advantage of one of the main strengths of CartograPlant: the curation and integration of a diversity of data types (genotypic, phenotypic and environmental) from different studies, thanks to the metadata collection using ontologies and standards.

**CONCLUSIONS**
CONCLUSIONS

• **CartograPlant** is intended to serve as a community resource for Plant Molecular Ecology.

• These flexible analytic workflows allow to analyze a diversity of data types (e.g. SNPs, SSR) and experimental designs (e.g. natural populations, common garden) and facilitate a diversity of biological questions in CartograPlant.

• Mega and meta-analysis take advantage of one of the main strengths of CartograPlant: the curation and integration of a diversity of data types (genotypic, phenotypic and environmental) from different studies, thanks to the metadata collection using ontologies and standards.

• **CartograPlant** is continuously growing to stay up-to-date with the latest research and statistical methods to analyze high-throughput biological data.
CONCLUSIONS

- **CartograPlant** is intended to serve as a community resource for Plant Molecular Ecology.

- These flexible analytic workflows allow to analyze a diversity of data types (e.g. SNPs, SSR) and experimental designs (e.g. natural populations, common garden) and facilitate a diversity of biological questions in CartograPlant.

- Mega and meta-analysis take advantage of one of the main strengths of CartograPlant: the curation and integration of a diversity of data types (genotypic, phenotypic and environmental) from different studies, thanks to the metadata collection using ontologies and standards.

- CartograPlant is continuously growing to stay up-to-date with the latest research and statistical methods to analyze high-throughput biological data.

- Having a centralized and up-to-date platform to integrate, visualize and analyze high-throughput biological data is key in the current big data era in plant biology.
Members of the project

- Stephen P. Ficklin
- Nic Herndon
- Emily Grau
- Sean Buehler
- Shay Muhonen
- Risharde Ramnath
- Umed Singh
- Charles Demurjian
- Meghan Myles
- Emily Strickland
- Victoria Burton
- Maddie Gadomski
- Jill Wegrzyn
- Margaret Staton
- Abdullah Almsaeed

This project has been funded by USDA-NIFA #2018-09223

https://cartograplant.org/