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
Big Data in Genomics

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.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Byte</td>
<td>1</td>
</tr>
<tr>
<td>Kilobyte</td>
<td>1,000</td>
</tr>
<tr>
<td>Megabyte</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Gigabyte</td>
<td>1,000,000,000</td>
</tr>
<tr>
<td>Terabyte</td>
<td>1,000,000,000,000</td>
</tr>
<tr>
<td>Petabyte</td>
<td>1,000,000,000,000,000</td>
</tr>
<tr>
<td>Exabyte</td>
<td>1,000,000,000,000,000,000</td>
</tr>
<tr>
<td>Zettabyte</td>
<td>1,000,000,000,000,000,000,000</td>
</tr>
</tbody>
</table>

Mostly Genomic but... Proteomics, Phenomics

NCBI Sequence Read Archive (SRA)

21,063,423,032,084,288 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
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 Observe

Watching plants and recording observations is easy. Start by selecting a plant that interests you or one that you observe in your neighborhood. For each observation, include the date, location, and specific plant you observed. Observations help us understand how plants are responding to climate change and other environmental factors.

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

Challenges of Integratingomics: ‘Big Data’ explained by Data Science

Volume
40 Zettabytes (40,000,000,000,000,000,000 bytes) of data will be created by 2020, an increase of 300 times from 2005.

6 Billion People have cell phones
World population: 7 Billion

It’s estimated that 2.5 Quintillion Bytes (2.5 x 10^18 bytes) of data are created each day.

Most companies in the U.S. have at least 100 Terabytes (100,000 Gigabytes) of data stored.

The New York Stock Exchange captures 1 TB of trade information during each trading session.

Variety
420 Million Wearable, wireless health monitors

30 Billion pieces of content are shared on Facebook every month

As of 2011, the global size of data in healthcare was estimated to be 150 Exabytes (150 Billion Gigabytes).

By 2014, it’s anticipated there will be 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

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

By 2015, it is projected there will be 18.9 Billion network connections – almost 2.5 connections per person on earth

Analysis of streaming data

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 increase 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.

1 in 3 business leaders don’t trust the information they use to make decisions

27% of respondents in one survey were unsure of how much of their data was inaccurate

Veracity
1.3 Trillion a Year

Poor data quality costs the US economy around $3.1 Trillion a Year.
Challenges of Integratomics

Forest Tree Research Community Survey

- Computational Resources
- Hosting Data on the Web
- Integrating Data Across Databases

Bar chart showing the challenges of Integratomics, with 'Computational Resources' being the highest challenge.
Improving infrastructure for tree genomic and phenomic data
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
3,100 unique visitors from 116 countries
Tripal Framework in TreeGenes

- Plant PopGen Submit (TPPS)
- JBrowse
- CartograTree

- Web
- Database: CHADO + Web Services
- Analysis: Galaxy
Tripal Gateway Project: Bringing Analytical Capacity to the Data

Chado + Drupal

Challenge #1

Data Exchange

Challenge #2

Data Transfer

Challenge #3

Data Analysis
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)

Population Study
- Publication
- Species

Study Design
- Landscape
- Common Garden
- Greenhouse
- Growth Chamber
- Breeding (Plot)

Phenotype, Genotype, Environment
- Georeferenced

Raw Data
Minimal Information About a Plant Phenotyping Experiment (MIAPPE)

MIAPPE

Minimum Information About a Plant Phenotyping Experiment

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

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:

Ontology Integration:

- Plant Ontology (PO)
- Crop Ontology (CO)
- Trait Ontology (TO)
- Environmental Ontology (ENVO)
- Chemical Ontology (CHEBI)
Phenotype Metadata File: Please upload a file containing columns with the name, attribute, description, and units of each of your phenotypes.

File: phenotype metadata.xlsx

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:

<table>
<thead>
<tr>
<th>Phenotype Name/Identifier</th>
<th>Attribute</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>phenotype 1</td>
<td>age</td>
<td>years</td>
<td>quantitative</td>
</tr>
<tr>
<td>phenotype 2</td>
<td>age</td>
<td>years</td>
<td>quantitative</td>
</tr>
<tr>
<td>phenotype 3</td>
<td>age</td>
<td>years</td>
<td>quantitative</td>
</tr>
</tbody>
</table>

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

Phenotype Description:

Please provide a short description of Phenotype 2

Phenotype Units:

- Humidity regime

[Map data ©2018 Google, in Terms of Use]
# TreeGenes to CartograTree

## TPPS/TGDR DETAILS FOR TGDR001

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession</td>
<td>TGDR001</td>
</tr>
<tr>
<td>Title</td>
<td>Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (Populus trichocarpa, Salicaceae) secondary xylem.</td>
</tr>
<tr>
<td>Species</td>
<td>Populus trichocarpa</td>
</tr>
<tr>
<td>Study Type</td>
<td>GxP</td>
</tr>
<tr>
<td>File Downloads</td>
<td>✓ ASSOCIATION RESULTS FILE 1 FILE</td>
</tr>
<tr>
<td></td>
<td>✓ GENOTYPES SNP</td>
</tr>
<tr>
<td></td>
<td>✓ GPS COORDINATES</td>
</tr>
<tr>
<td></td>
<td>✓ HAPLOTYPE DATA FILE</td>
</tr>
<tr>
<td></td>
<td>✓ PHENOTYPES</td>
</tr>
<tr>
<td></td>
<td>✓ PHENOTYPES DEFINITIONS</td>
</tr>
<tr>
<td>CartograTree</td>
<td><strong>View in CartograTree</strong></td>
</tr>
<tr>
<td>Tree Count</td>
<td>1376</td>
</tr>
<tr>
<td>Phenotype Count</td>
<td>4032</td>
</tr>
<tr>
<td>Unique Phenotypes</td>
<td>3</td>
</tr>
</tbody>
</table>

Plant Computational Genomics Lab | Department of Ecology & Evolutionary Biology | UConn
CartograTree: Integrating environmental layers with georeferenced trees
CartograTree: Integrating environmental layers with georeferenced trees
CartograTree: Integrating environmental layers with georeferenced trees
CartograTree: Save searches locally and select for meta-analysis
Bringing Analytical Capacity to the Data

Tree Genes

Hardwood Genomics

Genome Database for Rosaceae

CartograTree
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

Galaxy Training Network

Tweets by @galaxyproject

Galaxy Project
@galaxyproject

Galaxy Platform News: EGI_einfra ChemFlow, ChIP-Seq Docker, @usegalaxyp and @GalaxyAustralia galaxyproject.org/news/2018-01-g...
Galaxy: Open Source Web-based Platform for Bioinformatic Analysis
Workflows for Landscape Genomics: Integrating across diverse datasets

Genotypic data

Galaxy

Correlated markers

Environmental data
Workflows: Executed in Galaxy with metadata

Pre-processing

Reduce dimensionality

Select G x E data

Filter by G x E data

Filter by SNPs

Filter by MAF

Filter by individuals

Filter by MGF

Population structure

Landscape genetics analysis

Display results

Heterozygosity
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