The topological shape of gene expression across the evolution of flowering plants

Sourabh Palande[1] Sarah Percival[1] Aman Husbands[2] Arjun Krishnan[1] Beronda Montgomery[1] Elizabeth Munch[1] Addie Thompson[1] Alejandra Rougon-Cardoso[3] Daniel Chitwood[1] Robert VanBuren[1] Class of Fall 2020 HRT-841[1]

[1]Michigan State University

[2]University of Pennsylvania

[3]UNAM, ENES-Lyon

November 3, 2022

Acknowledgement

This work was made possible by the hard work of **HRT-841: Plants & Python** students!

Joshua Kaste, Miles Roberts, Kenia Segura Abá, Carly Claucherty, Jamell Dacon, Rei Doko, Thilani Jayakody, Hannah Jeffery, Nathan Kelly, Andriana Manousidaki, Hannah Parks, Emily Roggenkamp, Ally Schumacher, Jiaxin Yang, Jeremy Pardo.

Interactive "Intro to Python" JupyterBook:
https://plantsandpython.github.io/PlantsAndPython/

Introduction

Motivation

Grand Challenge: Relating genotype to phenotype: development, environment, evolution, ...

- **Gene expression:** Common currency across scales.
- **Molecular level:** DNA, -omics. Responsible for gene expression complexity.
- **Organism level:** Cell-specific expression, development orchestrated by gene expression.
- **Population level:** Life history, evolution.
- **Ecological level:** Climate, Global distribution of species, etc.

Outline

Grand Challenge: Relating genotype to phenotype: evolution, development, environment.

- 900 mil. years of evolution, Over 300K gene expression data sets.
- Typically, 1-specie or 1-gene studies.
- **Goal:** A meta-study of gene expression across all species.
- **Challenge:** Tremendous biological complexity, data set heterogeneity.
- **Approach:** Reduce heterogeneity, use Mapper for visualization.
- **Observations:**
  - Core, conserved backbone defining plant form and function.
  - Patterns differentiating plant tissues, biotic and abiotic stresses.

Data

Data Collection

- Selected 16 plant families, 54 distinct species.
  - Broad phylogenetic diversity within angiosperms.
  - High quality reference genome.
  - Breadth of tissue and stress types.
- Sample expression, metadata: NCBI BioProject, SRA, primary publications.
- Raw RNAseq data processed through common analytical pipeline.
- $\approx$ 3200 samples, 2671 left after processing.
- 8 tissue types.
- 9 biotic and abiotic stresses (+ healthy samples!)

Frequency Plots

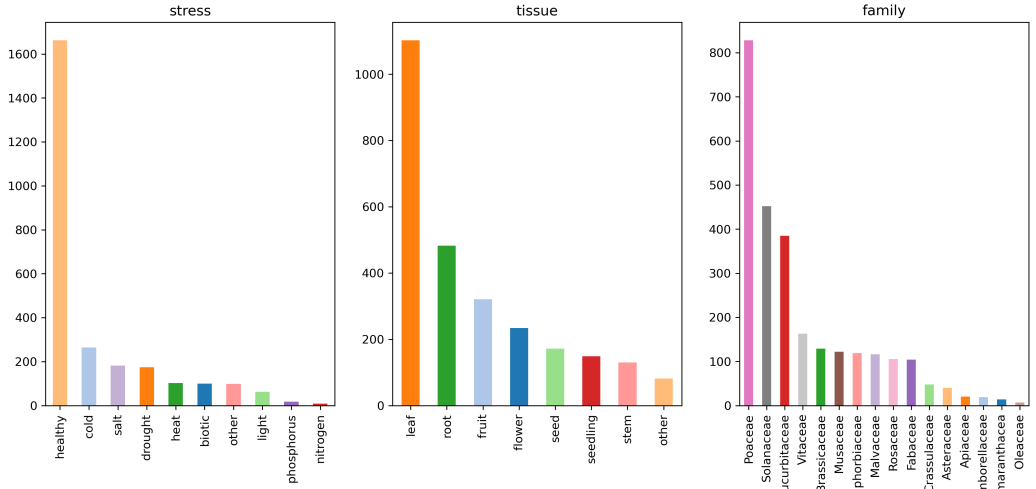

Figure: Factor Frequency Plots

Finding Orthogroups

- Cross-species comparisons: Need correspondences.
- *Orthogroups:* Groups of genes with similar function across species.
- Orthofinder: sequence alignment and clustering.
- Excluded multi-gene families with diverse functions.
- Excluded genes with high copy number.
- 2 million genes $\rightarrow$ 6328 orthogroups.
- TPM counts summed for genes in an orthogroup.
- Highly diverse, heterogeneous data combined into single expression matrix.

Dimension Reduction 1

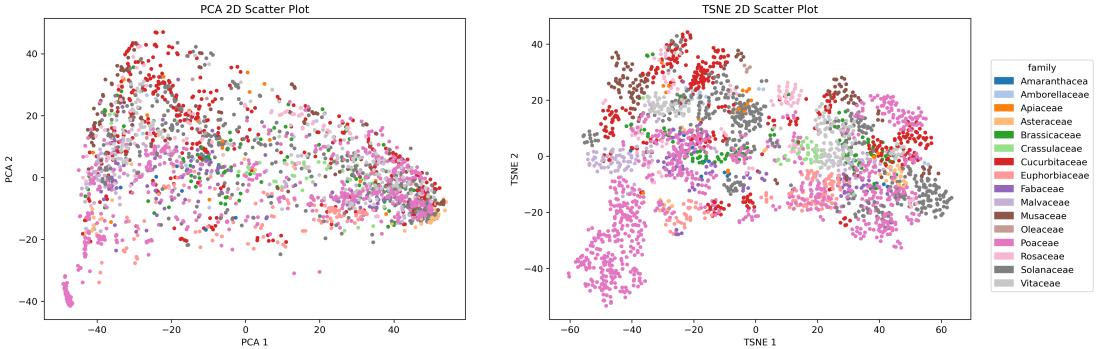

Figure: Dimension Reduction: Points colored by Family.

Dimension Reduction 2

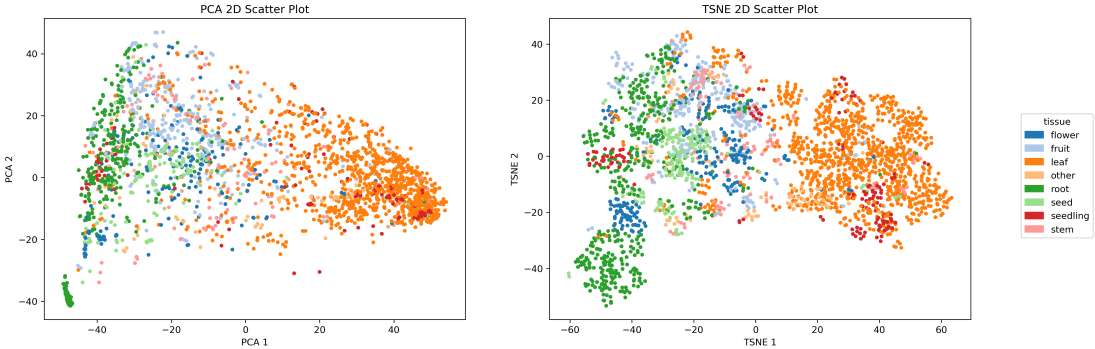

Figure: Dimension Reduction: Points colored by Tissue type.

Dimension Reduction 3

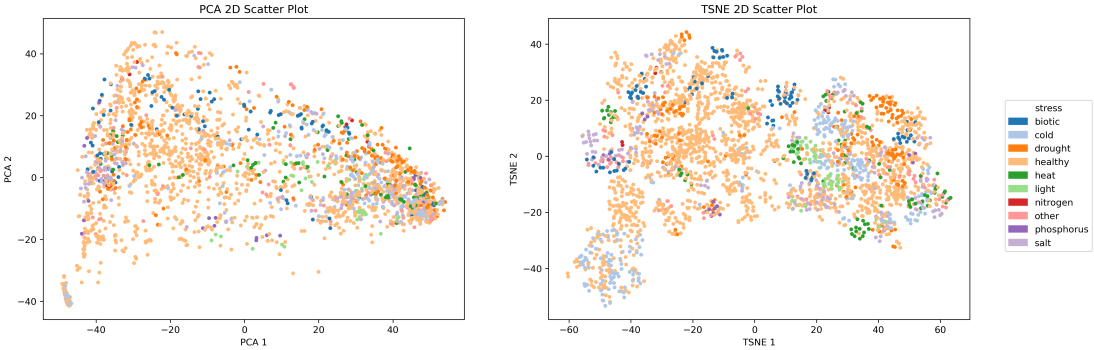

Figure: Dimension Reduction: Points colored by Stress type.

Mapper

Mapper Algorithm

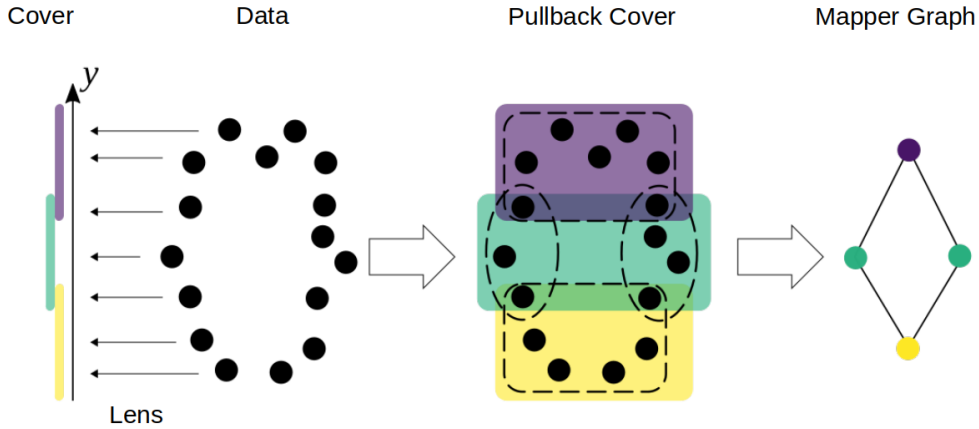

Figure: Mapper Algorithm

Mapper: Key Components

- Choice of lens: Domain / application dependent.
  - Only observe structure visible through specified lens.
  - Induce priors, domain knowledge.
  - We'll focus on creating good lens functions.
- Choice of cover:
  - Determines connectivity, density of output graph.
  - Heuristics for optimal cover choice available.
  - Usually - trial and error.
- Clustering algorithm:
  - Pick your favorite!
  - We stick to the default: DBSCAN.

Creating Lenses

- Two lenses[a]: Tissue lens, Stress lens.
- Pick a base class: *healthy* vs *stressed*, *leaf* vs *other*.
- Fit a linear model: *ideal* expression for base class.
- Project all samples on to the linear model.
- Residuals: Deviation from *ideal* expression.
- Use norm of the residual as lens.

[a] Nicolau, Levine, and Carlsson 2011.

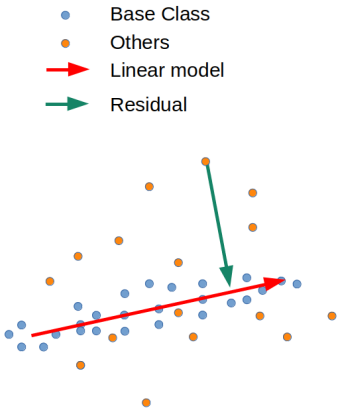

Figure: Creating lens

Lens Correlation Analysis

- For a given lens:
- For each orthogroup:
  - Compute Lens-Orthogroup correlation.
- 6328 correlation values.
- 2.5% most +ve correlations (right tail).
- 2.5% most -ve correlations (left tail).
- 159 Orthogroups in each tail.
- GO Enrichment Analysis for subset vs all.

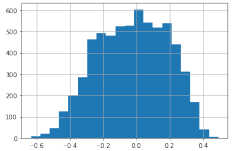

Figure: Leaf lens

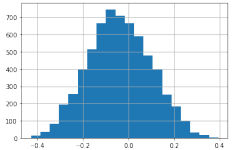

Figure: Stress lens

GO Enrichment

- Gene Ontology (GO): Standardized vocabulary describing gene function.
  - Cellular component.
  - Molecular function.
  - Biological Process.
- Enrichment: test subset of genes vs all.
  - Does the subset contain more representative of certain GO terms compared to chance?
  - Fischer's exact / Hyper-geometric test.
- We want to test a subset of orthogroups vs all.
- GO terms only available for genes, not orthogroups!

GO Enrichment

- Use Arabidopsis genome as reference.
  - Also used to find orthogroups.
  - Why?: Model organism, very well studied genome.
- Use orthogroup - Arabidopsis gene correspondence.
- Create orthogroup - GO term associations.
- GO Analysis tools: https://pypi.org/project/goatools/
- Perform enrichment analysis for each tail separately.
- Correct for multiple testing.

Results

Go Enrichment Results

- Tissue lens: Captures photosynthetic vs non-photosynthetic divide.
- GO enrichment of +ve correlated orthogroups:
  - Core metabolic processes, development of non-photosynthetic tissues.
- GO enrichment of -ve correlated orthogroups:
  - Related to photosynthesis, response to light, chloroplast organization.
- Stress lens: healthy vs stressed gene expression
- GO enrichment of +ve correlated orthogroups:
  - Genes involved in stress response.
- GO enrichment of -ve correlated orthogroups:
  - Genes involved in growth and reproduction.

Mapper: Tissue Lens

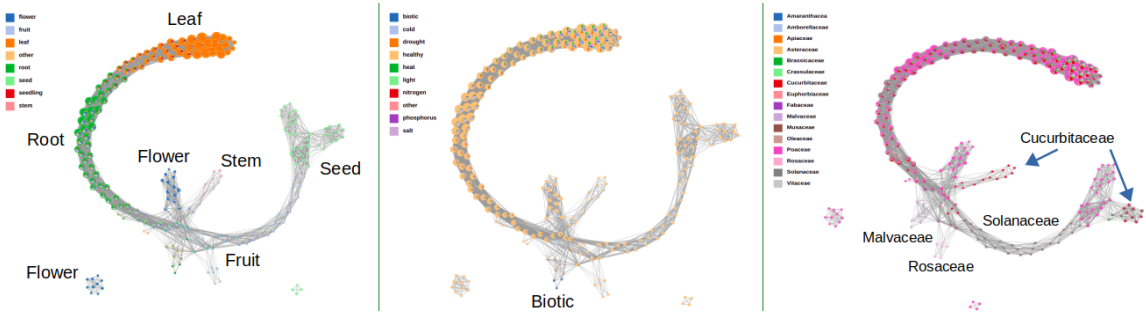

Figure: Tissue (leaf) Mapper Visualization

Mapper: Stress Lens

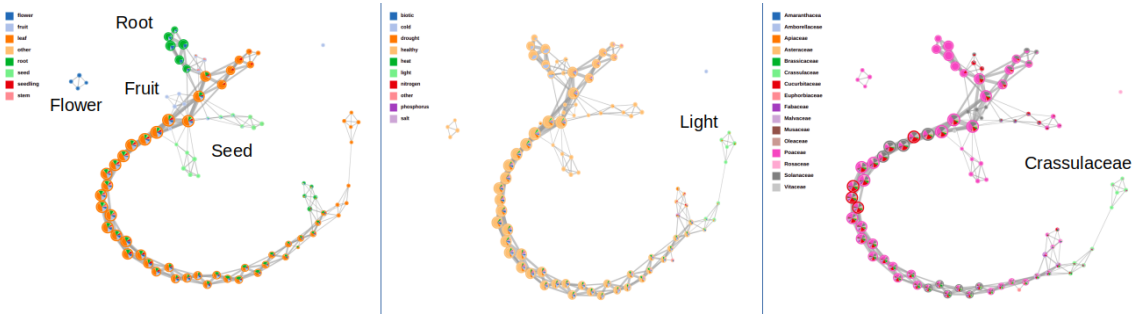

Figure: Stress Mapper Visualization

Discussion

Summary

- First cross-species expression study.
- Curated a novel data set.
- Gene expression data observed through phenotype lenses.
- Tissue lens: captures life cycle of plants.
- Stress lens: Deeply conserved stress response signatures.

Future Work

- Proof of concept: Cross-species gene expression can be valuable.
- Expand to include the wealth of public gene expression data sets
- Go beyond mapper. More sophisticated genome alignment and analysis methods.
- Transfer learning across species?

Preprint: https://www.biorxiv.org/content/10.1101/2022.09.07.506951v1

Data and Code: https://github.com/PlantsAndPython/plant-evo-mapper

email: palandes@msu.edu

References I

Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival". In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 7265–7270. DOI: 10.1073/pnas.1102826108.

Mapper: Root Lens

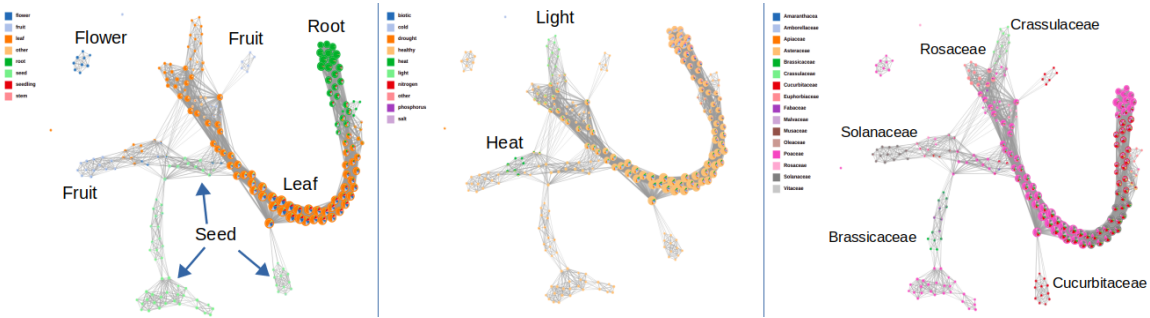

Figure: Root Mapper Visualization