PhytoOracle: Modular, Scalable Phenomic Data Processing Pipelines

Emmanuel Gonzalez*[1] (emmanuelgonzalez@email.arizona.edu), Ariyan Zarei,[2] Nathanial Hendler,[1] Michele Cosi,[1] Jeffrey Demieville,[1] Travis Simmons,[1,3] Holly Ellingson,[4] Nirav Merchant,[4] Eric Lyons,[1,4] Duke Pauli,[1,4] and **Andrea Eveland**[5]

[1]School of Plant Sciences, University of Arizona, Tucson, AZ; [2]Department of Computer Sciences, University of Arizona, Tucson, AZ; [3]College of Coastal Georgia, Brunswick, GA; [4]Data Science Institute, University of Arizona, Tucson, USA; and [5]Donald Danforth Plant Science Center, St. Louis, MO

PhytoOracle GitHub Repository, CyVerse Data Commons

Project goals

As phenomics continues to generate larger and higher dimensional data sets, there is an urgent need to develop and implement robust data processing pipelines to extract biological insight and knowledge from these data. Current phenomics processing pipelines lack modularity and the ability to exploit distributed computational infrastructure. To address these challenges, we developed PhytoOracle (PO), a suite of modular, scalable pipelines that aim to improve data processing efficiency for plant science research. PO integrates CCTools' Makeflow and Workqueue frameworks for distributed task management on local, Cloud, or high-performance computing (HPC) systems. Each pipeline component is available as a standalone container, providing transferability, extensibility, and reproducibility. PO efficiently extracts phenotype trait values, capturing phenotypic variation for elucidation of the genetic components of complex traits.

Abstract

Understanding how plants dynamically respond to environmental conditions has been historically difficult due to the low throughput and long-term cost of longitudinal data collection in the field setting[1]. Recent technological advances have resulted in small, cheap, and high resolution sensors that can be used to rapidly collect phenotypic trait data at regular time intervals in field or greenhouse settings[2,3]. Such platforms will continue evolving and become more widespread, likely exacerbating the phenotyping bottleneck by creating a strain on current data processing frameworks[4].

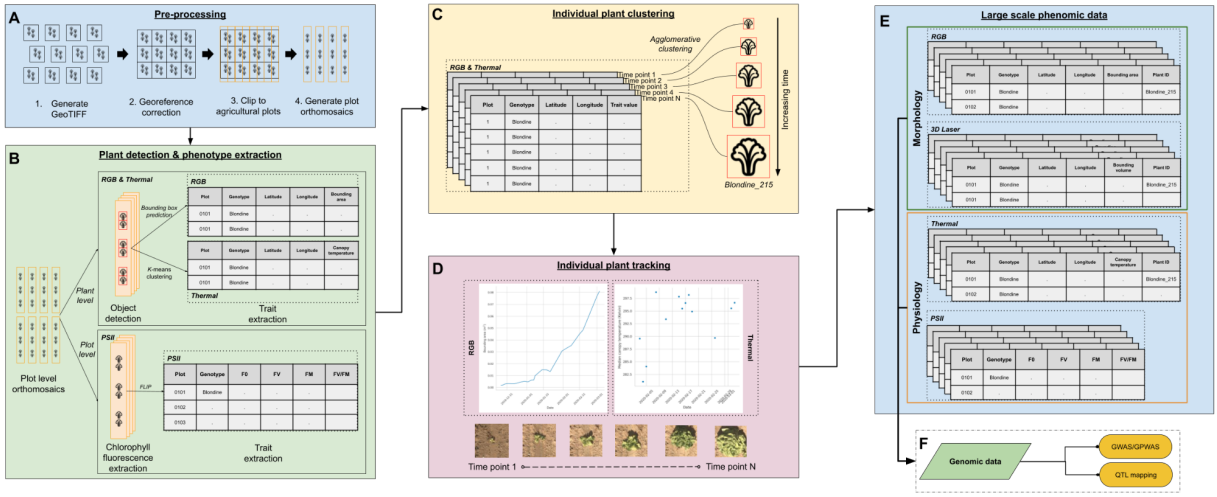

Figure 1 PhytoOracle pipeline workflow. (A) Image georeferencing metadata is corrected and images are clipped to agricultural plot boundaries. (B) Plants are detected using a Faster R-CNN model and (C) clustered and given a unique ID. (D) Single plants can be tracked using unique IDs. (E) Multimodal data is merged using unique IDs and plot information. (F) Large scale phenomic data and genomic data can enable GWAS and/or QTL mapping.

To address this phenotyping bottleneck, we developed PO, a suite of modular, scalable data processing pipelines for RGB and thermal cameras, Photosystem II (PSII) imager, and structured-light laser scanners (3D laser) raw data[5]. PO distributes tasks using CCTools Makeflow and Workqueue[6] to extract plant-level bounding area, canopy temperature, height, bounding volume and plot-level maximum quantum efficiency of PSII (F_V/F_M) using Faster R-CNN models (**Fig. 1**). PO scales with processing times of 235 minutes for 9,270 RGB images, 235 minutes for 9,270 FLIR images, and 13 minutes for 39,678 PSII

images (**Fig. 2**). The resulting processed data are associated using agglomerative clustering to enable time-series analysis of individual plant phenotypes (**Fig. 1**). Images and point clouds of each plant throughout the growing season are collected. These datasets are useful for a wide variety of future ML applications, such as the classification of plant varieties for precision agriculture, segmentation of plant disease, and modeling of plant growth and its influence on light use efficiency.

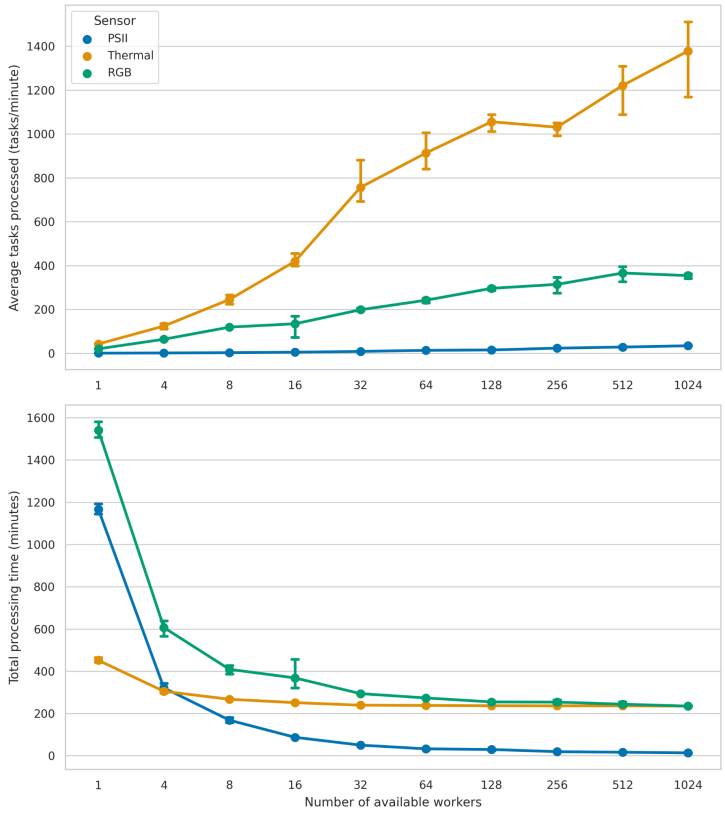

Figure 2 PhytoOracle benchmarking of PSII, thermal, and RGB pipelines. Increase in average tasks processed as the number of workers increases (Top). Decrease in total processing time (minutes) as the number of workers increases (Bottom). Benchmark datasets for each sensor were processed over the following range of available workers: 1, 4, 8, 16, 32, 64, 128, 256, 512, and 1024. Each configuration was run three times totaling 30 benchmark observations.

Funding statement

This material is based upon work supported by the U.S. Department of Energy Biological and Environmental Research program under Award Number DE-SC0020401, the U.S. Department of Energy Advanced Research Projects Agency - Energy OPEN program under Award Number DE-AR0001101, and the National Science Foundation CyVerse project under Award Number DBI-1743442.

References

1. Reynolds, D. *et al.* What is cost-efficient phenotyping? Optimizing costs for different scenarios. *Plant Sci.* **282**, 14–22 (2019).
2. Li, B. *et al.* Phenomics-based GWAS analysis reveals the genetic architecture for drought resistance in cotton. *Plant Biotechnol. J.* **18**, 2533–2544 (2020).
3. Sooriyapathirana, S. D. S. S. *et al.* Photosynthetic Phenomics of Field- and Greenhouse-Grown Amaranths vs. Sensory and Species Delimits. *Plant Phenomics* **2021**, 1–13 (2021).
4. Furbank, R. T. & Tester, M. Phenomics - technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* **16**, 635–644 (2011).
5. Gonzalez, E. *et al. PhytoOracle: Scalable, Modular Phenomic Data Processing Pipelines*. (2021).
6. Albrecht, M., Donnelly, P., Bui, P. & Thain, D. *Makeflow: a portable abstraction for data intensive computing on clusters, clouds, and grids*. 13 (2012).