PhytoOracle: Leveraging Open-Source Tools for Phenomic Data Processing at Scale

Michele Cosi[1*] (cosi@email.arizona.edu), Emmanuel Gonzalez,[1] Ariyan Zarei,[2] Travis Simmons,[1,3] Duke Pauli,[1] Eric Lyons,[1] and **Andrea Eveland**[4]

[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]Donald Danforth Plant Science Center, St. Louis, MO

https://github.com/LyonsLab/PhytoOracle
https://github.com/phytooracle
https://phytooracle.readthedocs.io/

Project Goals: Within the last decade, phenomics research has seen larger and higher-dimensional data sets. These data sets exceed the capacity for standard computational and analytical frameworks requiring the development of workflows capable of handling these vast amounts of data. To address these challenges, we developed PhytoOracle, aimed at improving processing and analysis of phenomics data. PhytoOracle speeds up analyses by leveraging distributed computing resources, and improves data reproducibility through containerization of computational code. PhytoOracle's distributive abilities and containerized code allow for handling of larger data volumes and modalities making it customizable and scalable. As a result, PhytoOracle can halve the time required to process 1TB of data, accelerating the overall extraction of morphological and physiological parameters. Ultimately, derived phenotypes are quantified and associated to causal genetic components in order to study abiotic stress tolerance.

Abstract: Plant phenomics is the scientific field that aims to quantify and study phenotypic traits through the application of sensor technology and machine learning algorithms. As sensor technology advances, data volume and processing times increase. Due to the lack of open source phenomic pipelines, our team developed PhytoOracle, a modular, scalable pipeline that aims to improve analysis of phenomics data. Our pipeline addresses the issues through (1) accelerating analysis tasks by integrating distributed computing resources and managing high throughput data; (2) containerization of computational code for improved ease-of-use and reproducibility. PhytoOracle expedites data processing by distributing tasks to either local, cloud, or high-performance computing (HPC) systems using CCTools (Albrecht, Donnelly, Bui, & Thain, 2012). Pipeline components are available as Docker containers, providing portability and modularity. Containerized code allows users to execute code without the need to install additional software dependencies, proving to be an efficient solution for deployment. PhytoOracle was developed for, but is not limited to, processing data primarily originating from the world's biggest agricultural robot, the Scanalyzer, located at the University of Arizona's Maricopa Agricultural Center. The Scanalyzer is equipped with sensors that are able to capture a variety of sensor data at sub-millimeter resolution, outputting up to 10 terabytes (TB) of data per

day. These sensors include a laser line scanner, a hyperspectral imager, and three cameras: thermal infrared, RGB, and chlorophyll fluorescence. The capacity of data generated by the Scanalyzer easily surpasses the processing capacity of standard laboratory computers, therefore requiring a quicker data processing solution. PhytoOracle is able to process 1TB of data in half the time a 64-core laboratory computer requires, by distributing jobs across 2,700 HPC cores. The advancement of phenomics demands algorithms capable of processing increasingly large data volumes within a reasonable timeframe. As a result of these key capabilities, PhytoOracle can efficiently process data in a timely manner to extract phenotypic information, which in turn enables faster elucidation of the genetic components of complex traits.

This work is supported by grants DOE-BER #DE-SC00248484, DOE ARPA-E #DE-AR0001101.

References:

Albrecht, M., Donnelly, P., Bui, P., & Thain, D. (2012). Makeflow: A portable abstraction for data intensive computing on clusters, clouds, and grids. SWEET ’12, 1–13. doi:10. 1145/2443416.2443417