# PhytoOracle Phenomics Data Processing Pipelines
## Preparing for a New Season

Emmanuel Miguel Gonzalez

2023-12-20

# Table of Contents

# 1 Field Preparation

The following steps must be completed prior to planting:

1. Shape raised beds
2. Set up sprinkler irrigation pipes, heads, gaskets, and filters
3. Inject subsurface drip irrigation tape
4. Place string and labeled stakes in the field

   *Note: These steps are carried out by Pauli lab members a few weeks before planting.*

After completing these steps, the field will look like **Figure 1**.

***Figure 1:*** *South gantry field with shaped raised beds, sprinkler irrigation, and strings and stakes.*

## 1.1 Planting

Lettuce planting generally occurs around Mid-November to early-December. The mean air temperature during this time has previously ranged from 10 °C to 22 °C (**Figure 2**).

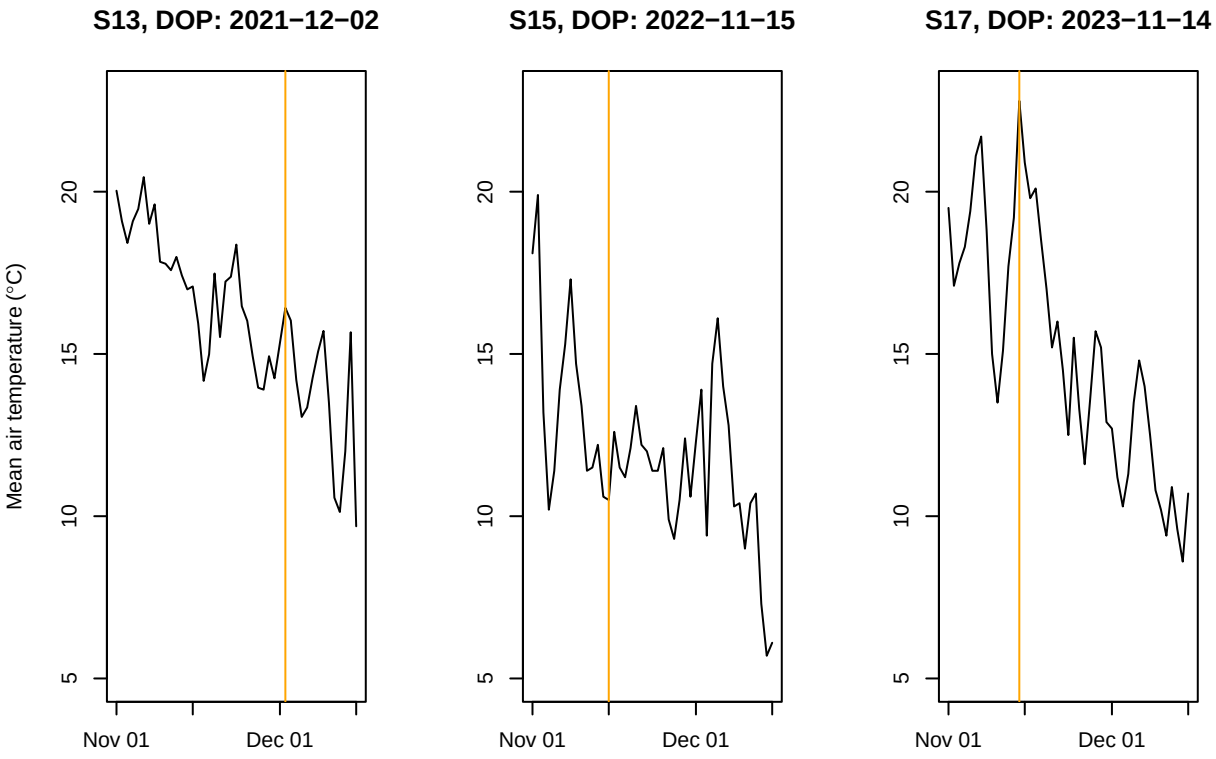


***Figure 2:*** *Mean air temperature data collected by The Arizona Meteorologial Network (AZMET). Orange vertical lines represent the day of planting. DOP, day of planting; S13, season 13; S15, season 15; S17, season 17.*

### 1.1.1 Equipment

Planting is done by hand using Earthway garden seed planters (**Figure 3 Left**). Lettuce seeds must be planted at a depth of 1/8 to 1/4 inch. The planting depth can be set using the adjustable screw at the bottom of the seed planter - ensure this is set to an acceptable depth throughout planting as it can shift. Also, make sure that the chain is not tangled at the bottom of the planter, as it is meant to cover the soil after the planter penetrates the soil during planting. If the chain is tangled, seeds will not be covered with soil, and thus, may not germinate or be blown/washed away.

The Earthway planters were modified by fitting them with funnels and tubing that allows the user to manually hand-feed the small lettuce seeds instead of using the provided seed container and plates. Planting is carried out by members of the Pauli, Arnold, and Michelmore labs. People are paired up with one person responsible for planting the seeds with the Earthway planter, and the other responsible for ensuring the correct plot numbers are being planted and that the correct seed is provided to the person planting (**Figure 3 Right**).

***Figure 3:*** *Lettuce hand planting. (Left) Earthway garden seed planter. (Right) One person planting using the Earthway planter, while the other is responsible for ensuring correct plot numbers and handing the correct seed to the person planting.*

### 1.1.2 Potential Issues During Planting

In past years, the tubing that feeds the seeds into the ground have gotten pinched or otherwise clogged. In these cases, entire columns were inadequately planted - the seed did not make it into the seed line of the expected plot. When this happens, Drs. Duke Pauli and Maria

José Truco are notified. The plots within the specific column/s are noted. If seed is not immediately available, Dr. Maria José Truco sends it from Davis, California.

## 1.2 Ground Control Points

The raw data collected by the Field Scanalyzer has a high level of misalignment of images and point clouds. To mitigate this error, a high number of ground control points (GCPs) are placed in the field. These GCPs include (**Figure 4**): - White plastic bucket lids, four columns into the field on both east and west ends - Umbrella holders with grey metal bucket lids, trench between four and five columns into the field on both east and west ends

***Figure 4:*** *Ground control points (GCPs) used in the gantry field. (Left) White plastic bucket lid. (Right) Umbrella holder with grey metal bucket lid.*

Each range contains a single white plastic bucket lid and two umbrella holders with grey metal bucket lids in the following arrangement (**Figure 5**):

## 1.3 Thinning

Thinning is a very important part of the field trial. The planters often result in clusters of seeds germinating close to each other. Thinning is conducted in two phases (**Figure 6**):

- Phase 1: Thin the plots into 10 plant clusters
- Phase 2: Thin the plots into 10 individual plants

The 10 individual plants resulting from Phase 2 should be equidistant. The equidistant placement reduces any overlap with neighboring plants. This is an important step as the goal with the Field Scanalyzer data is to phenotype each plant individually. The farther plants are, the easier it is to individually phenotype them.

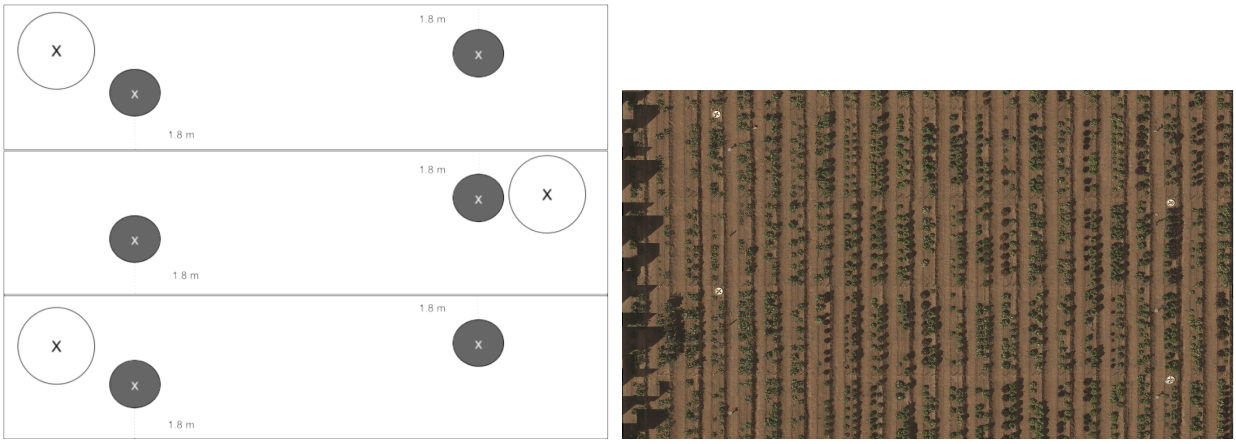


***Figure 5:*** *Arrangement of ground control points (GCPs) in the gantry field. (Left) Each range contains a single white plastic bucket lid and two umbrella holders with grey metal bucket lids. (Right) White plastic bucket lids are alternated, to ensure robust geocorrection.*

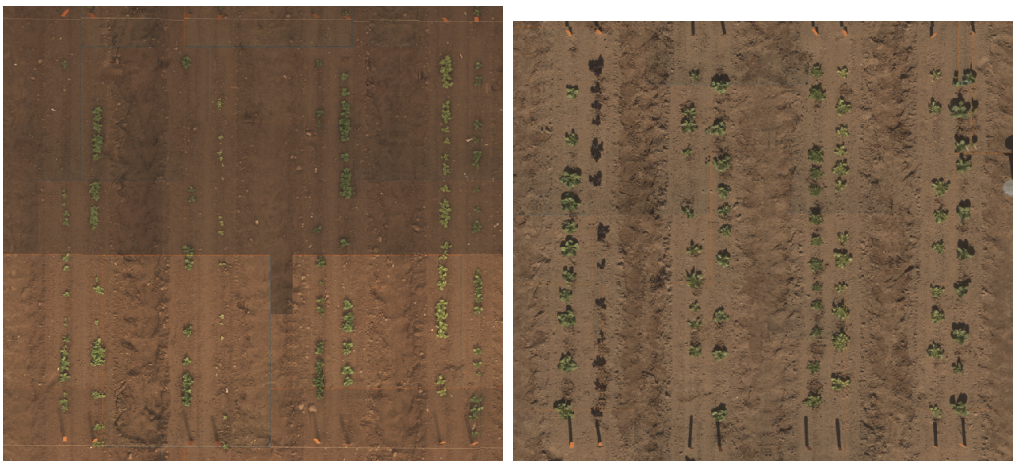

***Figure 6:*** *Change in plant density after multiple rounds of thinning. (Left) Plants after Phase 1 of thinning. (Right) Plants after Phase 2 of thinning.*

# 2 Positioning Information Preparation

The Global Positioning System (GPS) coordinates of each GCP must be collected so they can be used in PhytoOracle workflows. To accomplish this, you need a Trimble Global Navigation Satellite System (GNSS) (**Figure 7**).


***Figure 7:*** *Trimble Global Navigation Satellite System (GNSS) used to collect accurate Global Positioning System (GPS) coordinates of Ground Control Points (GCPs).*

## 2.1 Collecting Global Positioning System (GPS) Coordinates

The United States Department of Agriculture (USDA) Arid Land Agricultural Research Center (ALARC) has trimbles that we can borrow. To use them, follow the steps below:

1. Run Trimble Access - Press Trimble hard key (Windows symbol), select Trimble Access
2. Log in — Click either “Tap here to log in” or the current logged in person (e.g., kelly.thorp)

- Login Mode: Offline
- Select your user name.
  - If you already have a username, select it. . . click next. . . click finish
  - Else type in your user name. . . Click finish
  - Passwords are optional. . . be professional

3. Set up a job - Click General Survey -> Jobs

- If opening existing job, click “Open job” then select the job
- If starting a new job, click “New job” then
  - Job name: Give the job a name.
  - Template: Default
  - Coord sys.: Click ‘Select from library’
    - System: ‘World wide/UTM’
    - Zone: 12 Nonh
    - Datum: WGS 1984 (7P)
    - Project height: 361m
  - Click Store Accept
- If using the Yuma 2 and the Bluetooth needs connected
  - Turn on the GNSS receiver
  - General Survey -> Instrument -> GNSS Functions Bluetooth
    - Under ‘Connect to GNSS Rover’ select: R8s, Serial#: Trimble
    - Save/Accept

4. To measure points

- Go to field. Click General Survey -> Measure -> ALARCRTK -> Measure points
- Accept base station.
- Point name: Name the point. (If you put a number at the end, it will auto-increment.)
- Code: Add optional additional info.
- Method: Rapid point
- Antenna height (Uncorr): 2.000m
- Vertical offset: 30.839m Up (This is important to get the Z coordinate correct.)
- Click “Measure” to record a point.
- Go to lab. Connect TSC3 to computer.
- Click General Survey -> Jobs -> Import/Export -> Export fixed format -> Accept -> All points
- Copy the file off the TSC3:
  - “This PC\Trimble Navigation Limited TSC3\\Export”

5. To stake flags at point locations

- Copy CSV file to the TSC3
  - Save CSV from Excel to help TSC3 recognize points. Do not add header line.
  - Order: Point name, Northing, Easting, Elevation, Code (optional)
  - Connect TSC3 to computer and copy point file to your user directory.
    - “This PC\Trimble Navigation Limited TSC3\”
- Go to field. Click General Survey -> Stakeout -> ALARCRTK -> Points - Remove points from old stakeouts, if they exist. - Add -> Select from file -> choose point file. - Click ‘All’ - Click Add - Select a point, then select StakeoUt Navigate to the point, and stake flag.

***Table 1:*** *Ground Control Point (GCP) coordinate file. Each row represents the coordinate of a single GCP.*

| GCP | Type | Northing | Easting | Height..m. |
|---|---|---|---|---|
| plate1 | White | 3659979 | 408992.8 | 360.775 |
| plate2 | White | 3659987 | 408992.9 | 360.788 |
| plate3 | White | 3659995 | 408992.9 | 360.783 |
| plate4 | White | 3660003 | 408993.0 | 360.770 |
| plate5 | White | 3660011 | 408993.1 | 360.775 |
| plate6 | White | 3660019 | 408993.1 | 360.765 |

***Table 2:*** *Ground Control Point (GCP) coordinate file. Each row represents the coordinate of a single GCP.*

| GCP number | Latitude | Longitude |
|---|---|---|
| 1 | 33.07470 | -111.975 |
| 2 | 33.07478 | -111.975 |
| 3 | 33.07485 | -111.975 |
| 4 | 33.07492 | -111.975 |
| 5 | 33.07499 | -111.975 |
| 6 | 33.07506 | -111.975 |

## 2.2 Positioning Information Files Required by PhytoOracle

PhytoOracle relies on geospatial information, such as GPS coordinates, to accurately link phenotypes with a location in the field. This allows us to detect, tag, and track individual plants over the course of multiple Field Scanalyzer scans. Specifically, PhytoOracle requires two files:

1. GCP file: Text file containing the GPS coordinates of all field GCPs.
2. GeoJSON: File containing polygons representing each plot in the gantry field

These files must be generated prior to data processing for the respective season. Additionally, these files should be loaded onto QGIS for visual inspection and confirmation that the coordinates are accurate.

### 2.2.1 Generating GCP File

The Trimble collects GPS coordinates in the Easting, Northing format (Table 1). PhytoOracle requires GPS coordinates to be in the latitude, longitude format. To convert the coordinates, click here to use the conversion tool. After running the conversion script, the data will now be in the required latitude, longitude format (Table 2).

### 2.2.2 Generating GeoJSON File

GeoJSON files contain polygons that represent each plot in the gantry field (**Figure 8**). These polygons are used to extract smaller experimental units from larger units, such as the full field scale.

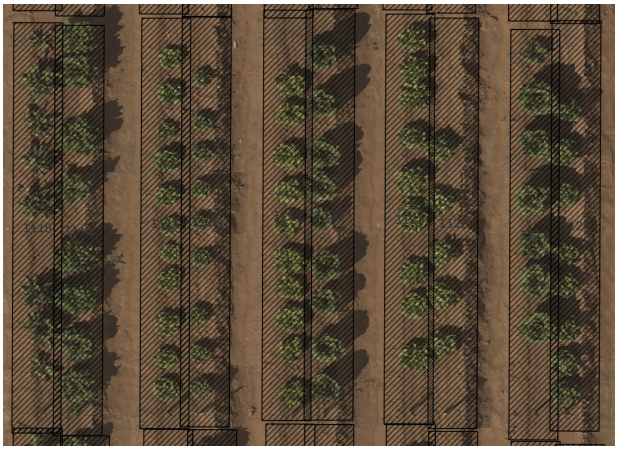

***Figure 8:*** *GeoJSON file containing a single polygon for each plot.*

Our field design and dimensions remain pretty consistent from one season to the next. As a result, existing GeoJSONs are modified and applied to new seasons. In the case that a new GeoJSON needs to be created, please refer to FIELDimageR.

If you are editing a pre-existing GeoJSON, you will need to:

1. Move polygons that are misaligned in the new season
2. Rename genotype column

**2.2.2.1 Moving polygons** To move polygons, you need to load the GeoJSON and a drone orthomosaic onto QGIS. Then, you can follow the steps in **Figure 9**:

1. Click "Toggle Editing"
2. Click "Select Features by Area or Single Click"
3. Click "Move Features"
4. Manually move polygon into desired alignment
5. Single click to drop the polygon into the desired location
6. Save changes

**2.2.2.2 Renaming genotype column** The "genotype" values in the GeoJSON file can be edited using GeoPandas. A GeoJSON can be opened up as a dataframe, similar to Pandas. Once opened, you can then replace the "genotype" columns using the fieldbook for the respective season. To see an example click here.

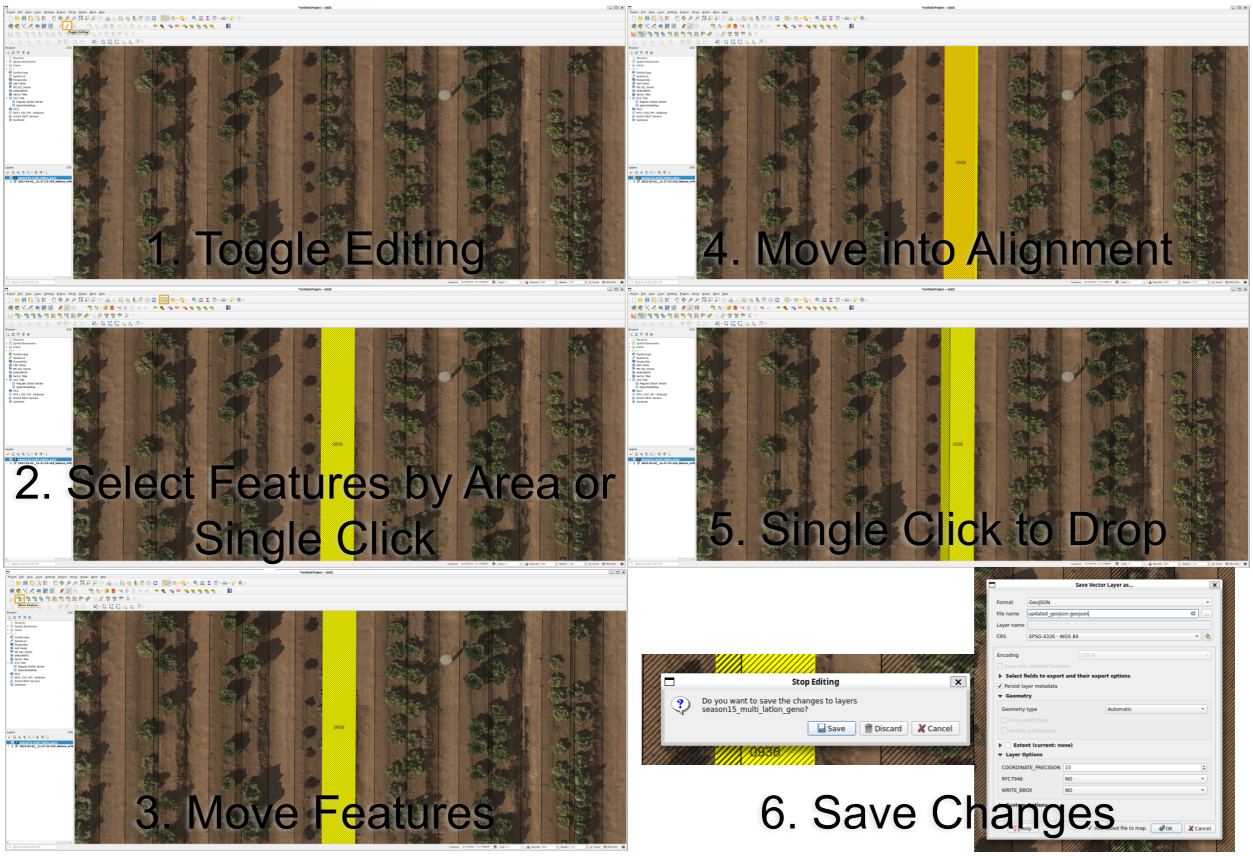


***Figure 9:*** *Editing GeoJSON polygons using QGIS.*

# 3 Pipeline Preparation

The PhytoOracle (PO) pipelines require the aforementioned GCP and GeoJSON files. Additionally, a Yet Another Markup Language (YAML) file is used by PO for automated, reproducible data processing. YAMLs are a form of a configuration file that can be used to define a series of arguments/flags. The details of the YAML files can be found on our PhytoOracle Automation repo: click here.

## 3.1 Editing YAML file

For each season, YAML files must be edited to correctly process data for the respective season.

Specifically, the following keys should be edited for each season:

- tags
  - pipeline
  - season
  - season_name
- workload_manager
  - manager_name
- paths
  - cyverse
    - input
      - necessary_files

Examples of YAMLs for each season can be found here.

## 3.2 Updating GitHub Repositories

### 3.2.1 PhytoOracle Data

At the start of a new season, the phytooracle_automation repo. Specifically, the season_config_yaml variable needs to be updated: click here.

The season is defined by multiple keys, including name, start_date, end_date, flir_temp_units, and complete_field_dates (**Figure 10**).

Below are some details for each key:

- name: Season name, as it shows up on the CyVerse DataStore

- start_date: Date of planting
- end_date: Date of harvest
- flir_temp_units: Temperature units reported by thermal (FLIR) images
- complete_field_dates: List of RGB scan dates

```
181     13:
182         name: season_13_lettuce_yr_2022
183         start_date: 2022-11-14
184         end_date: 2023-03-30
185         flir_temp_units: "C"
186         complete_field_dates:
187                 rgb:
188                     ['2022-01-27__10-54-27-164_lettuce',
189                     '2022-01-31__10-45-39-463_lettuce',
190                     '2022-02-01__10-39-38-575_lettuce',
191                     '2022-02-04__10-42-09-085_lettuce',
192                     '2022-02-07__10-33-28-355_lettuce',
193                     '2022-02-08__10-38-14-413_lettuce',
194                     '2022-02-09__10-42-41-102_lettuce',
195                     '2022-02-10__10-37-53-213_lettuce',
196                     '2022-02-11__10-37-09-800_lettuce',
197                     '2022-02-14__10-41-53-763_lettuce',
198                     '2022-02-16__10-43-57-766_lettuce',
199                     '2022-02-17__10-39-08-670_lettuce',
200                     '2022-02-18__10-41-21-294_lettuce',
201                     '2022-02-21__10-38-28-308_lettuce',
202                     '2022-02-25__10-39-57-210_lettuce',
203                     '2022-02-28__10-38-14-633_lettuce',
204                     '2022-03-02__10-56-11-405_lettuce',
205                     '2022-03-07__10-38-37-481_lettuce',
206                     '2022-03-09__10-40-08-384_lettuce',
207                     '2022-03-11__10-39-22-200_lettuce',
208                     '2022-03-15__11-12-27-310_lettuce']
209
```

***Figure 10:*** *Section of the season$_c$onfig$_y$aml variable in the phytooracle$_d$ata GitHub repository.*

To get a list of RGB dates, use iRODS to ils the directory of the respective season (**Figure 11**).

### 3.2.2 PhytoOracle Landmark Selection

The PhytoOracle 3d_landmark_selection contains the phytooracle_data repository. As such, the 3d_landmark_selection container must be rebuilt once the abovementioned changes have been made to phytooracle_data repository. To do so, access the container here and click on “Trigger” for the “latest” container (**Figure 12**).

```
eg@myosotis: .../ubuntu_files
eg@myosotis: ... /ubuntu_files$ ils /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/ | grep
 2022
/iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce:
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-01-27__10-54-27-164_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-01-31__10-45-39-463_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-01__10-39-38-575_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-04__10-42-09-085_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-07__10-33-28-355_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-08__10-38-14-413_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-09__10-42-41-102_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-10__10-37-53-213_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-11__10-37-09-800_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-14__10-41-53-763_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-16__10-43-57-766_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-17__10-39-08-670_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-18__10-41-21-294_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-21__10-38-28-308_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-25__10-39-57-210_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-02-28__10-38-14-633_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-03-02__10-56-11-405_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-03-07__10-38-37-481_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-03-09__10-40-08-384_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-03-11__10-39-22-200_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2022-03-15__11-12-27-310_lettuce
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2222-01-01__11-24-18-691_sorghum
  C- /iplant/home/shared/phytooracle/season_13_lettuce_yr_2022/level_1/stereoTop/lettuce/2222-02-02__00-00-00-000_lettuce
```

***Figure 11:*** *Getting RGB dates using iRODS.*

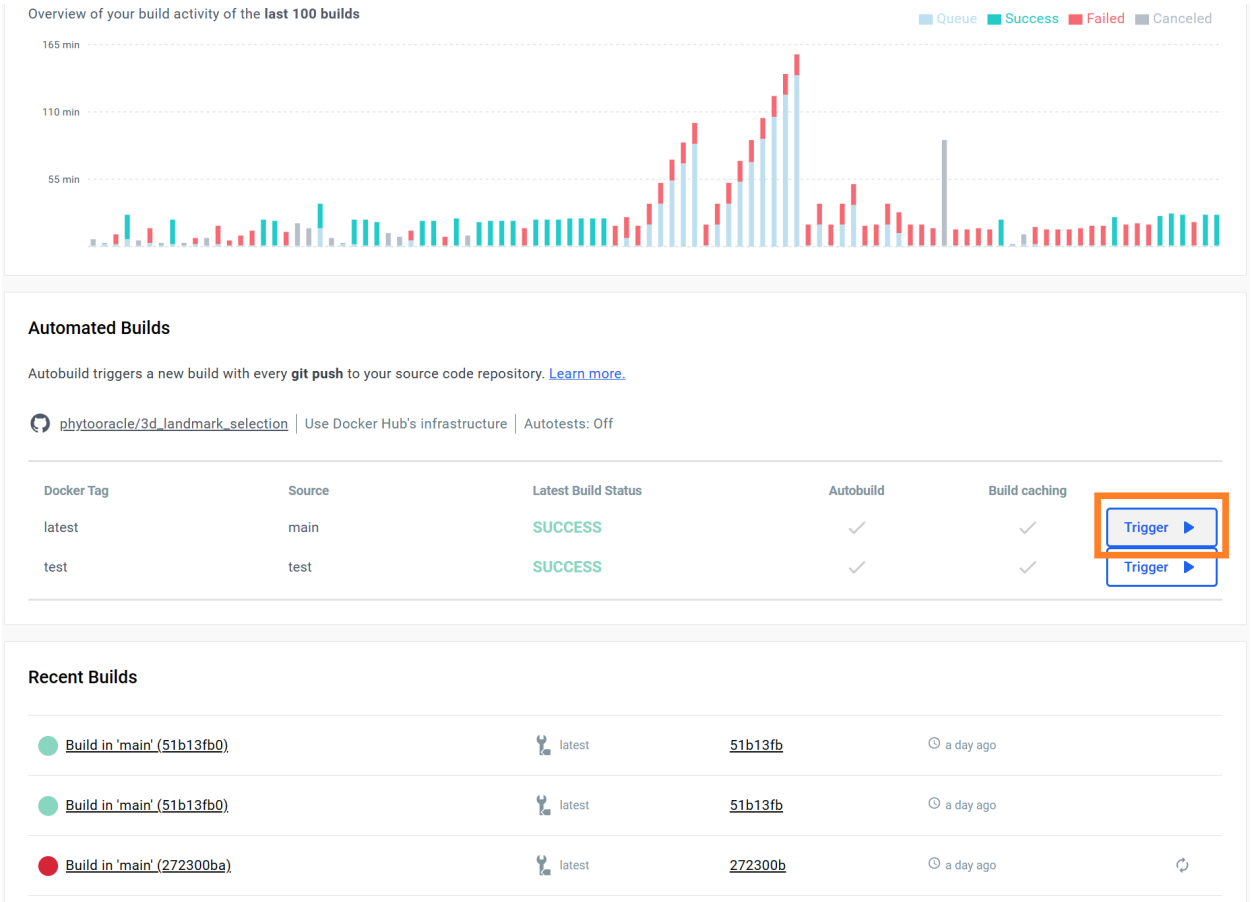


***Figure 12:*** *Rebuilding the* $3d_landmark_selection$ *container on DockerHub.*

# 4 Intro to High Performance Computers

## 4.1 UArizona High Performance Computing Cluster

The University of Arizona maintains an HPC center, which houses three compute resources: El Gato, Ocelote, and Puma.

### 4.1.1 Compute System

| Name | El Gato | Ocelote | Puma |
|---|---|---|---|
| Model | IBM System X iDataPlex dx360 M4 | Lenovo NeXtScale nx360 M5 | Penguin Altus XE2242 |
| Year Purchased | 2013 | 2016 (2018 P100 nodes) | 2020 |
| Node Count | 131 | 400 | 236 CPU-only<br>8 GPU<br>2 High-memory |
| Total System Memory (TB) | 26TB | 82.6TB | 128TB |
| Processors | 2x Xeon E5-2650v2 8-core (Ivy Bridge) | 2x Xeon E5-2695v3 14-core (Haswell)<br>2x Xeon E5-2695v4 14-core (Broadwell)<br>4x Xeon E7-4850v2 12-core (Ivy Bridge) | 2x AMD EPYC 7642 48-core (Rome) |
| Cores / Node (schedulable) | 16c | 28c (48c - High-memory node) | 94c |
| Total Cores | 2160* | 11528* | 23616* |
| Processor Speed | 2.66GHz | 2.3GHz (2.4GHz - Broadwell CPUs) | 2.4GHz |
| Memory / Node | 256GB - GPU nodes<br>64GB - CPU-only nodes | 192GB (2TB - High-memory node) | 512GB (3TB - High-memory nodes) |
| Accelerators | | 46 NVIDIA P100 (16GB) | 29 NVIDIA V100S |
| /tmp | ~840 GB spinning /tmp is part of root filesystem | ~840 GB spinning /tmp is part of root filesystem | ~1440 TB NVMe /tmp |
| HPL Rmax (TFlop/s) | 46 | 382 | |
| OS | Centos 7 | CentOS 7 | CentOS 7 |

| Name | El Gato | Ocelote | Puma |
|---|---|---|---|
| Interconnect | FDR Inifinband | FDR Infiniband for node-node<br>10 Gb Ethernet node-storage | 1x 25Gb/s Ethernet RDMA (RoCEv2)<br>1x 25Gb/s Ethernet to storage |

### 4.1.2 Compute Resources

The UArizona HPC provides three types of resources:

- Windfall: Unlimited, can be preempted
- Standard: Limited, no preemption
  - El Gato: 7,000 CPU-hours per month
  - Ocelote: 70,000 CPU-hours per month
  - Puma: 100,000 CPU-hours per month
- High Priority:
  - Puma
    - ericlyons: 175,200 CPU-hours per month
    - dukepauli: 35,040 CPU-hours per month

*Note: High priority is only available for the Puma cluster.

# 5 Running PhytoOracle on High Performance Computers

PhytoOracle is a scalable, modular phenomics data processing workflow manager. In short, this means that PhytoOracle can leverage high performance computer (HPC) clusters and cloud computing to distributed tasks across hundreds to thousands of cores.

## 5.1 Defining Compute Resources

Resources are defined in the workload_manager section of the PhytoOracle YAML. In this section, you can define the following settings:

- account: ericlyons
- high_priority_settings:
  - use: True

  - qos_group: user_qos_ericlyons
  - partition: high_priority

- standard_settings:

  - use: False
  - partition: standard

- job_name: phytooracle_worker_rgb
- nodes: 1
- number_worker_array: 490
- cores_per_worker: 1
- time_minutes: 720
- retries: 1
- port: 0
- #mem_per_core: 32
- mem_per_core: 5
- manager_name: stereoTop_level01_s15
- worker_timeout_seconds: 43200

## 5.2 Before Deploying

## 5.3