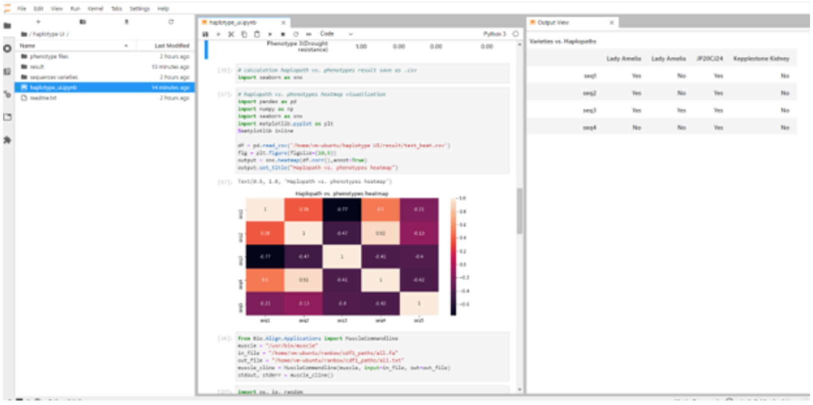
**HIP WP1.3: Molecular markers for maturity and yield**
**Report 2021**
**Project lead: WR, Plant Breeding Richard Visser; WU, Plant Breeding Sara Bergonzi**

In this project, we aim to investigate the system of molecular “taps” and “valves” in the plants’ plumbing system that facilitate the partitioning of resources to the potato tuber. We use this knowledge for the rationalized breeding of potato with optimized efficiency of carbon transport and positive knock-on effects on yield stability and stress tolerance. To achieve this, we will identify all variants of transporters and regulators in potato and develop a mathematical model of potato sugar transport and its regulation.

**Highlights.** After the establishment of the specification document for the web-based haplotyping tool, work began on creating a gold standard test dataset for validation of our approach. The *StSP3D* and *StSP6A* loci were previously used to prototype the visualization of haplotypes but focus was shifted to the more diverse *StCDF1* gene alleles. Using a combination of short and long read technologies and conventional haplotyping methods, the *StCDF1* locus was haplotyped for 6 diverse potato varieties. These haplotypes were verified with independent approaches, including genetic maps. *StCDF1* is a perfect test case as it includes a range of haplotypes, including haplotypes with a small 7 bp insertion and a large 865bp transposon insertion. These haplotypes have also shown strong correlations to different phenotypic outcomes for tuber development and abiotic stress resistance. In a connected project on the PAN Genome of potato the results of these analysis were recently published (Hoopes et al, 2022 Molecular Plant, DOI:https://doi.org/10.1016/j.molp.2022.01.003)

We developed an algorithm leveraging k-mers and de Bruijn graphs as well as pedigree information across a large dataset. The algorithm builds a graph of all the possible paths for a loci based on the k-mers of 150 potato varieties and then, based on the commonality of paths between varieties, coverage and the pedigree relationships of the varieties, reduces the paths to a set of the most likely “haplo-paths” for each variety. To enhance the usability of the algorithm, an application programming interface (API) was designed, which allows for easy interfacing. Furthermore, validation of the developed algorithm has begun using the gold standard dataset, with the 6 varieties, for the *StCDF1* locus and the “haplo-paths” generated by the algorithm for those same varieties among the set of 150 potato varieties.  Two other test cases were also developed. The disease resistance gene *RySto* provides an interesting case for discriminating between paralogs and orthologs. Similarly, *StSP5G* provides a unique use case with paralogs and orthologs but also variation between the number of exons in each gene copy. The ability of the algorithm to distinguish between these cases accurately will be key to its future success.

To facilitate easy inspection by the end-user, a prototype of an interactive webpage (based on Jupyter Notebook) with a variety of graphs and statistics to help users quickly and easily interpret the results has been developed (Figure 1). Jupyter Notebook is a standard notebook interface widely used by the data science community with expansive support that makes it both simple for the developers to develop and maintain in a reproducible manner, while also being familiar and straight forward enough for even non-bioinformaticians to use.



| Name | Last Modified |
|---|---|
| phenotype files | 2 hours ago |
| result | 13 minutes ago |
| sequences varieties | 2 hours ago |
| haplotype_ui.ipynb | 14 minutes ago |
| readme.txt | 2 hours ago |

| | | | | |
|---|---|---|---|---|
| Phenotype 3(Drought resistance) | 1.00 | 0.00 | 0.00 | 0.00 |

```
# calculation haplopath vs. phenotypes result save as .csv
import seaborn as sns
```

```
# haplopath vs. phenotypes heatmap visualization
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

df = pd.read_csv("/home/vm-ubuntu/haplotype UI/result/test_heat.csv")
fig = plt.figure(figsize=(10,5))
output = sns.heatmap(df.corr(),annot=True)
output.set_title("Haplopath vs. phenotypes heatmap")
```

```
Text(0.5, 1.0, 'Haplopath vs. phenotypes heatmap')
```

| | seq1 | seq2 | seq3 | seq4 | seq5 |
|---|---|---|---|---|---|
| seq1 | 1 | 0.16 | -0.77 | 0.5 | -0.21 |
| seq2 | 0.16 | 1 | -0.47 | 0.52 | -0.13 |
| seq3 | -0.77 | -0.47 | 1 | -0.41 | -0.4 |
| seq4 | 0.5 | 0.52 | -0.41 | 1 | -0.42 |
| seq5 | -0.21 | -0.13 | -0.4 | -0.42 | 1 |

```
from Bio.Align.Applications import MuscleCommandline
muscle = "/usr/bin/muscle"
in_file = "/home/vm-ubuntu/ranbow/cdf1_paths/all.fa"
out_file = "/home/vm-ubuntu/ranbow/cdf1_paths/all.txt"
muscle_cline = MuscleCommandline(muscle, input=in_file, out=out_file)
stdout, stderr = muscle_cline()
```

```
import os, io, random
```

Varieties vs. Haplopaths

| | Lady Amelia | Lady Amelia | JP20CJ24 | Kepplestone Kidney |
|---|---|---|---|---|
| seq1 | Yes | No | Yes | No |
| seq2 | Yes | No | Yes | No |
| seq3 | Yes | Yes | Yes | No |
| seq4 | No | No | Yes | No |

*Figure 1. Screenshot of the current Jupyter Notebook prototype including data files (left), a table of phenotypic behavior for each variety (right) and a heatmap of the correlation between the "haplo-paths" and phenotypes (middle)*

**Knelpunten:**
The bioinformatician directly responsible for this work left in Autumn 2021 so this work was taken over by others. Meanwhile a group of new Bioinformaticians have been hired and started to work on different projects including this one.

**Planning:**
In 2022, work will continue on developing, validating and applying the algorithm to ever larger datasets. For the *StCDF1* locus on chromosome 5, a comprehensive catalogue of haplotypes across a larger region that includes the CIS-acting regions of the *StCDF1* promoter and the divergently expressed *StFLORE* transcript will be assembled from the 137 Atlas cultivars.  Further testing will be done to see if any other available biological information from our 137 potato variety dataset will be used to more accurately and easily identify the correct "haplo-paths".
For this, phenotypic information is currently being retrieved from several sources and organized in the Germinate database. We will also attempt to scale up the analysis in terms of the available k-mers and the size of the loci to be haplotyped, with the ultimate goal of being able to haplotype whole genomes. Work will also continue on the design of the visualization notebook in conjunction with the lab scientist. The focus here will be to expand on visualizations that make the data easier to interpret and expand support for the API as it develops. We plan to make prototypes of the API and visualization notebook available online to a wider userbase and then start an iterative process with the end-users in improving both the algorithm and visualization based on their feedback.

In the research performed by the experimental biologists in the context of the MAMY HIP project, interesting variation has been identified at the *StSP5G* locus. Preliminary analysis suggest that that locus is highly variable, with several different observed structural variants affecting possible gene function and including sub-functionalization. From a computational side we will analyse the locus in the available tetraploid sequenced varieties to gain insight into its variability and establish a link with the phenotype.

**Producten:**
**None**