HIP WP1.3: Molecular markers for maturity and yield
Project lead: WR, Plant Breeding Richard Finkers; WU, Plant Breeding Christian Bachem

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.

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.

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 novices to use.
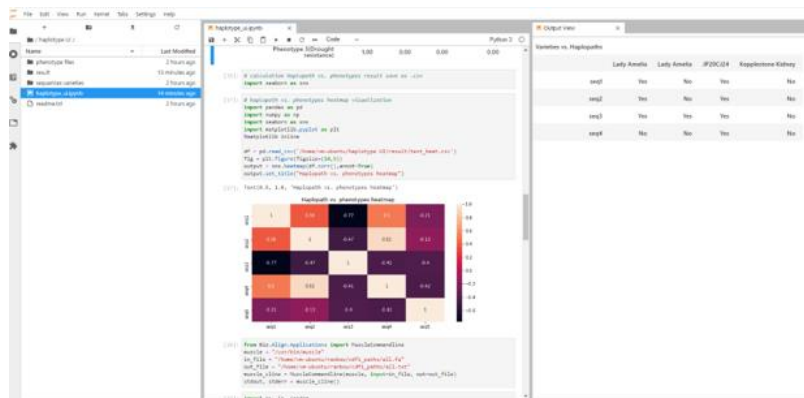

Figure 1. Screenshot of the current Juypter Notebook prototype including data files (left), a table of phenotypic behavior for each variety (right) and a heatmap of the correlation between the "haplo-paths" and phenotypes (middle)

Knelpunten:

None.

Planning:
In 2021, work will continue on developing, validating and applying the algorithm to ever larger datasets. In collaboration with the data scientist, validation will continue with the 6 haplotyped potato varieties for *StCDF1* and then expanded to other genes of interest such as *StSP3D* and *StSP6A.* Further testing will be done to see if any other available biological information from our 150 potato variety dataset can be used to more accurately and easily identify the correct "haplo-paths". 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.

Producten:
None.