This factsheet is a result of the first Wageningen University & Research (WUR) Data Science and AI Fellowship program. With this program we aim to increase and integrate our expertise in DS/AI throughout the entire organisation. The variety of projects highlights the potential for DS/AI across the WUR domains.

Author: Sven Warris

Machine Learning in plant pangenome research

Objective

We combine different data sets to analysis the genetic variability and its applicability for ML research of the *Capsicum* clade by creating pangenome information of over 100 plants.

Method

Gene expression data and the three Capsicum genomes were used for gene annotation. Genomic reads from nineteen-nine other peppers (pungent and non-pungent) were mapped against the three genomes for gene coverages and BUSCO/resistance genes (NLR) predictions. The pangenome graph was built based on protein homology and linked to the functional annotations. Clustering and visualisation were done with R and pheatmap.

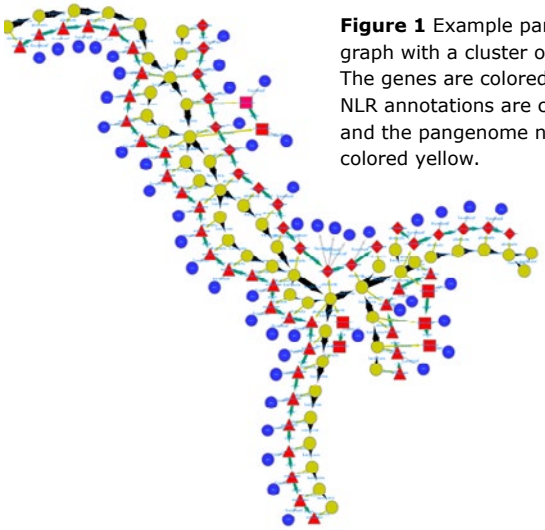

Figure 1 Example pangenome graph with a cluster of NLR genes. The genes are colored red, the NLR annotations are colored blue and the pangenome nodes are colored yellow.

Results

The resulting numbers of genes predicted are higher than expected when compared to the reference genomes. Genomes and annotations were added to a graph database and a pangenome graph was constructed (Figure 1). The BUSCO scores of the in-house assemblies are lower than the C. annuum public reference, but higher than the other references. High level of genetic diversity in NLR genes was observed in the gene coverage data. Accessions from the non-pungent cultivar Capsicum annuum miss many genes putatively related to the capsaicin pathway.

Impact

Through the functional annotations in the pangenome we identified genes putatively related to the capsaicin pathway or their regulatory genes and subsequently clustered. The results confirm the conserved nature of the BUSCO genes and the high-level genetic variability of the NLR genes. The non-pungent accessions differ from the pungent cultivars, however, the correlation to pungency of these presence/absence data do not follow directly.
The capsaicin pathway and its regulatory elements remain elusive. The size of the resequencing data set, combined with the pangenome plus annotations and the available phenotypic data of all the accessions provide a valuable data set for further machine learning analyses.

Future plans

For further research we would like to expand the data set with more phenotypic data. A follow-up project has already been started in collaboration with the new Netherlands Plant Eco-phenotyping Centre. The lessons learned for this project will be integrated into this new project as well as other upcoming projects where we combine pangenome information to phenotypic data.

Further information
The project ran from Janaury 2021 until Juli 2021.
Contact person: Sven Warris (sven.warris@wur.nl).

The scripts and data are in this repository:
https://git.wur.nl/warri004/ml-pangenomes

WAGENINGEN
UNIVERSITY & RESEARCH