PPP Annual Report 2019
PPP projects which are under supervision of the "Topsectoren" must report annually on the scientific content and financial progress. This form is used to report the progress of the content of the project. PPP projects that finish in 2019 should make use of a different form: "PPP-final report."
The annual report will be published on the TKI / topsector website. Therefore, please ensure that there is no confidential information in the annual report.
The PPP-annual report must be sent, at the latest, by the 1st of March 2020 to the "TKI's": info@tkitu.nl or info@tki-agrifood.nl. For Wageningen Research, the report has to be sent to the "Topsector secretary" of your respective institute.

General information	
PPP-number	KV 1605-118 TU-16003
Title	Building the Green hapmap
Theme	Food security and biotechnology
Implementing institute	Wageningen University and Research
Project leader research (name + e-mail address)	Elio Schijlen Elio.schijlen@wur.nl
Coordinator (on behalf of private partners)	Remco Ursem Remco.ursem@hzpc.nl
Project-website address	https://www.wur.nl/nl/Onderzoek-Resultaten/Onderzoeksprojecten-LNV/Expertisegebieden/kennisonline/TU-16003-Building-the-Green-HapMap.htm
Start date	02-01-2017
Final date	31-12-2020

Approval by the coordinator of the consortium The annual report must be discussed with the coordinator of the consortium. The "TKI's" appreciate additional comments concerning the annual report.	
Assessment of the report by the coordinator on behalf of the consortium:	☐ Approved ☐ Not approved
Additional comments concerning the annual report:	

Summary of the project	
Problem definition	In plant breeding, accurate haplotyping will have a major impact on genetic insight of many plant genomes include major food, ornamental and energy crops (e.g. potato, wheat) and thus finally on crop quality. 'Building the Green Hapmap' project opens new avenues for unlocking genetic diversity at the haplotype level and for facilitating optimal parent selection, to boost breeding in the Netherlands for efficient development of new global competitive varieties. This will not only address the ploidy challenge, but also decipher other unsolved issues with complex genomes: heterozygosity, structural variation, repeats and genome duplications.
Project goals	**T**he Green Hapmap project applies and adapts the newest DNA sequencing technology '10XGenomics' to obtain long range information from highest quality short read sequence data. The resulting 'linked-reads' span 10s to over 100 kilo bases and are used

	for alignment, variant analysis and reconstruction of phase blocks all at unprecedented levels. Within this project this technology will be specifically applied to and adapted for polyploid plant genome analysis. The project will focus on three plant species with increasing genome size, complexity and ploidy levels: diploid carrot, auto-tetraploid potato and auto-hexaploid chrysanthemum. In order to optimize the 10XGenomics linked read technologies for non-human genomes extensive adaptations of the current bioinformatics pipelines for haplotype phasing is required within this project.

Results

Planned results 2019

WP2: Optimizing 10X library preps and de-novo assemblies
WP3: Capture based complexity reduction for 10Xgenomics Linked read Sequence analysis and phasing
WP3: Cas9 based enrichment for 10Xgenomics Linked read Sequence analysis and phasing
WP4: Development of a bioinformatics pipeline that can be used for haplotype phasing of polyploid crops using the 10X Genomics technology

Achieved results 2019

WP2: Optimizing 10X library preps and de-novo assemblies.
Optimization of DNA mass, DNA fragment length and library combinations for 10X Genomics Linked reads Sequencing for phasing and de-novo assembly have been made. De-novo assembly of the heterozygous diploid carrot genome, the auto-tetraploid potato innovator genome and the auto hexaploid chrysanthemum genome were done using different adaptations and assemblers. The double haploid potato DM genome was used as control for comparison of the assembly results for the other genomes obtained by the supernova 2 assembler. Results showed that partially phased genome assemblies were obtained but also that genome assembly fragmentation rapidly increased with correlating genome ploidy and heterozygousity (table1)

Table1: statistics of 10XGenomics Linked redas bases de-novo assembly results

	Innovator	DM	Daucus	Chrysanthemum
assembler	Supernova 2	Supernova 2	Supernova 2	Supernova 2
Assembly Mb	1,378	746	628	5,154
LPM	ND	ND	ND	ND
Scfld	149,177	14,761	65,668	1,296,487
Scfld > 10Kb	22,728	2,023	14,388	84.1 K
N50 (Kb)	28	3,001	22	5.9
L50	5,220	66	5,320	192,258
Max (Mb)	6.0	16.3	1.7	1.0
Mol.size (Kb)	130	52	47	79




In addition, different assemblers (Supernova 2.0, Athena, ARCS and Megahit) were used to test their performance for the genome assembly of the tetraploid potato genome. The first three of these assemblers resulted in roughly comparable results whereas severe less assembly statistics were obtained from the Megahit assembly software. In conclusion our experience is that the Supernova assembler is performing best for de-novo assemblies and works well for diploid genomes with a relative low heterozygousity. Assembly of "true" 4n and 6n genomes appeared not feasible, instead collapsed diploid like versions were obtained.

WP3: Capture based complexity reduction for 10Xgenomics Linked read Sequence analysis and phasing.

Here we applied a bait based capturing method based on 1000's of probes targeting 14 specific regions of interest scattered all over potato genome (Selected TKI ATLAS gene regions, total 2Mb). In total 8 varieties were used for post 10X library prep enrichment and final sample specific 'enriched linked read sequencing' Reference based read mappings of linked sequence reads clearly showed successful enrichment of the target regions in all varieties ranging from 2.5 fold to nearly 12 fold overall enrichment (Table 2 and Figure 1).

Table2: Statistics of 10Xlinked reads Target Enrichment

	Anna	Claire	Fontana	Innovator	Melody	Novana	Seresta	Spunta
Mean coverage all targeted regions	**188**	**396**	**334**	**116**	**59**	**175**	**10**	**34**
Mean coverage non targeted regions	28	36	28	13	17	24	4	12
Enrichment (fold)	**6.7**	**11.0**	**11.9**	**8.9**	**3.5**	**7.3**	**2.5**	**2.8**
Data input (mbp)	8,883	15,586	14,246	10,000	4,506	6,731	1,017	2,456

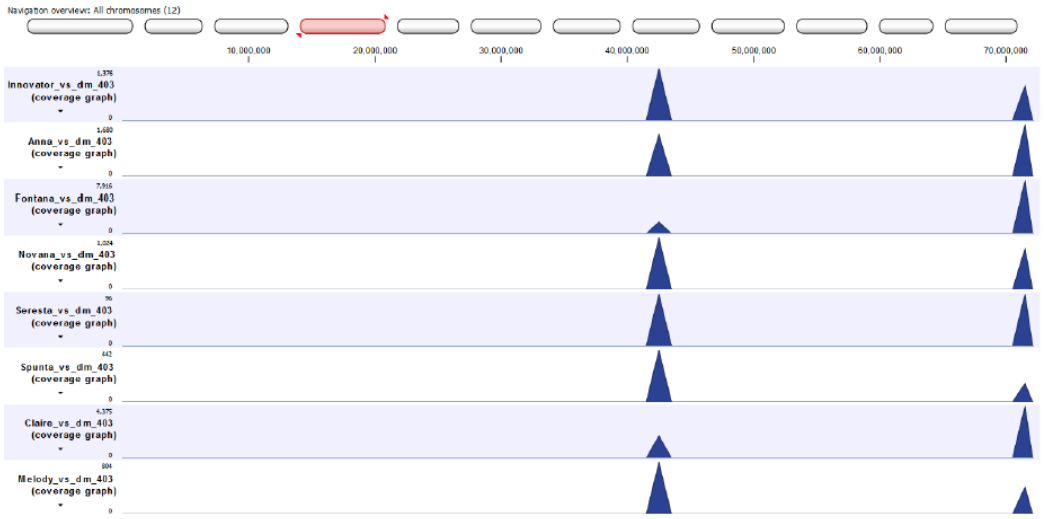

Figure 1. Coverage of 10Xgenomics linked reads showing two enriched regions of potato chromosome 4 across all eight varieties.

Data from above target regions was further used for validation of haplotype phasing results from the developed SDhaP pipeline (WP4). For this, 10X Linked Reads obtained from the enrichment of potato cv innovator were used for mapping and phasing with the long Ranger software (based on reference DM 4.03). Selected target region chr04 (position 42M) was further used as a representative region with one single Phase block spanning the complete enriched region. Mapped Linked reads were extracted from this region and used as input for the developed pipeline (adapted and running on HPC; mapping (bam from Long Ranger), Free Bayes variant calling, using tetraploid specific settings based on SNPs and filtering read depth (min 5 to max 500) with no repeat masking). Analysis using the SDhaP pipeline resulted for this region in 4 separated haploytpes. However visual inspection for this region indicated the existence of three real haplotypes whereas the 4th seemed to be collection bin of un-phased reads which seemed to be verified by biased read coverage (Fig 2). Therefore extensive validation and applying further quality filtering is absolutely required.
In conclusion, the bait based target enrichment worked well and resulting long molecules can be used for further haplotyping.

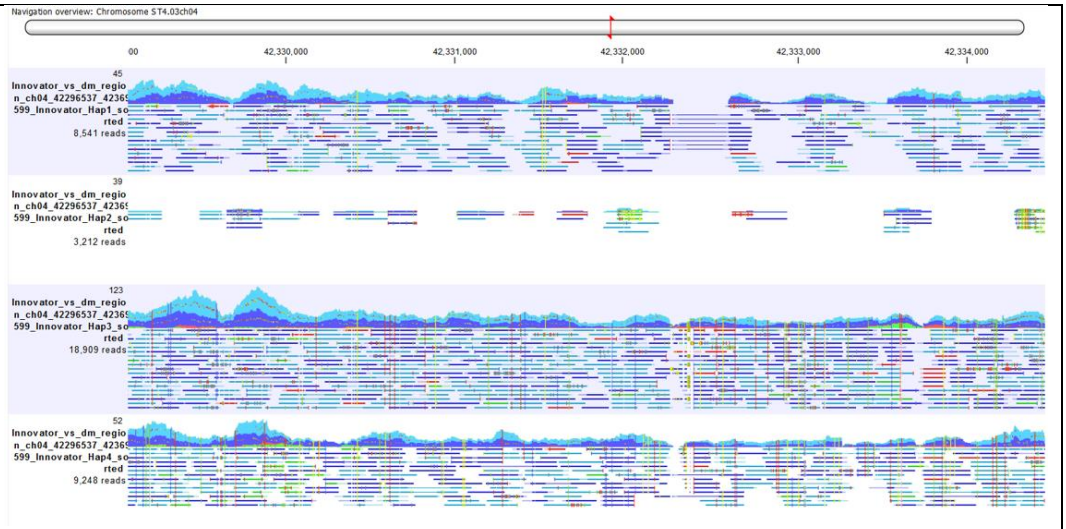

Figure 2. Mapping of linked reads from four reconstructed haplotypes from one region indicating three true haplotypes and one (low coverage) false haplotype.

WP4: Development of a bioinformatics pipeline that can be used for haplotype phasing of polyploid crops using the 10X Genomics technology.

The Haplotyping pipeline development consisted of; I) the 10x Genomics Long Ranger Pipeline for barcode processing and alignment. The Long Ranger pipeline is aborted and the alignment file (BAM file) is extracted. Based on the BAM file the read depth thresholds are calculated. For each defined region, the alignment data is extracted from the BAM file and this is used for variant calling using FreeBayes. The resulting VCF file is filtered and, based on the BAM file and the filtered VCF file, a SNP matrix is generated that serves as input for SDhaP for haplotype identification and reconstruction.

For validation of the haplotyping pipeline, 4 haplotypes were simulated based on chromosome 1 of potato DM v4.04. The haplotypes were generated by introducing random bi-allelic SNP's with a SNP rate of 1 per 46 bp. The dosage distribution of the SNPs over the 4 haplotypes is; 50% of the SNPs are simplex, 23% duplex, 14% triplex and 13% quadruplex. For the simulation, the same parameters as described in Ehsan Motazedi's publication, based on experimental data published by Uitdewilligen, were used. The simulated haplotypes were used to generate a 10XGenomics linked read dataset using LRSIM. This dataset was used as input for the haplotyping pipeline. For 5 defined regions (each 500 Kb) that were used in previous simulations and were selected to contain reference genes that were also used for validation in another haplotyping projects, the pipeline resulted in 4 haplotypes. For validation of the pipeline results, the haplotypes reconstructed were compared with the 4 simulated haplotypes as 'ground truth'. Based on all simulations, we conclude that the haplotyping pipeline is able to identify and reconstruct haplotypes for 500 kb regions. In case of regions with simulated reduced number of different haplotypes difficulties occur with haplotypes sharing the same alleles (sequence) for a large region. It looks like all conflicts are combined into the remaining haplotype, which is also indicated by Fig2.

In contrast to simulated data, experimental data would create difficulties to interpret the reconstructed haplotypes. Therefore an evaluation method to identify the reliability of the reconstructed haplotypes or haplotype regions is an absolute requirement in order to identify suspicious regions with erroneous haplotype reconstruction including haplotype shifts. Additional filtering including read depth was explored the reliability of haplotype reconstruction. As shown by figures 2 and 3 there were hardly any reads found in these cases supporting the fourth haplotype.

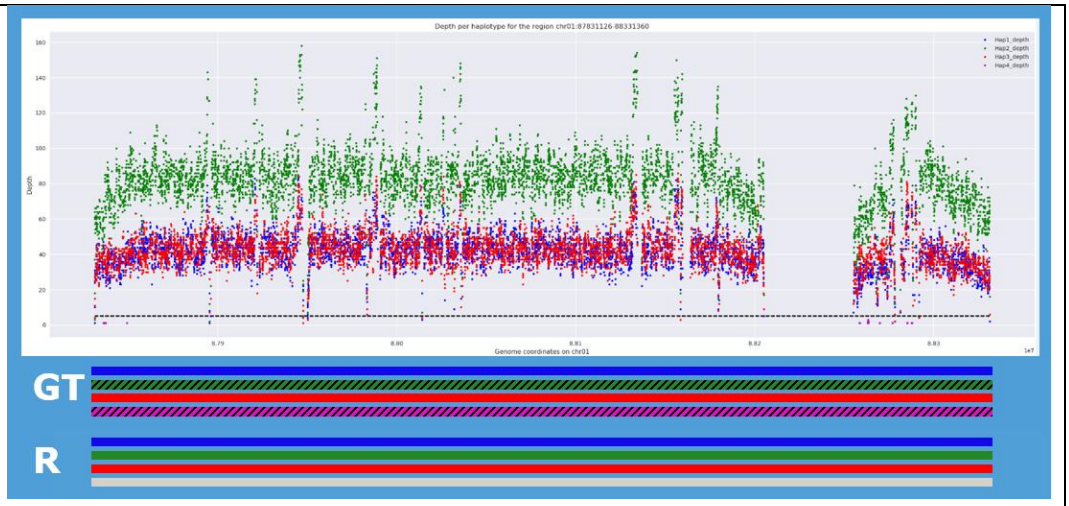

Fig 3. Read depth analysis of haplotype reconstruction (region 5) showing lack of support for the existence of a fourth haplotype missing in the reconstructed haplotypes (R, grey). The increased coverage for haplotype 2 (green) resulted from similar sequence reads from haplotype four.

Three methods for quality control were implemented; I) haplotype projection which helps to identify haplotypes that are the same II) VCF comparison which detects inconstancies between phased and un-phased VCF III) Molecule projection which indicates increased or reduced coverage and molecules switching between haplotypes. The three methods of the Quality control of the reconstructed haplotypes together appeared to be very useful for interpretation of reconstructed haplotypes. Unfortunately these methods are highly time and CPU memory consuming. Interpretation of the reconstructed haploytpes based on simulated data is clear, however interpretation of reconstructed haploytpes from experimental data is much more complex and requires careful further inspection.
As the haplotyping pipeline starts with barcode demultiplexing and alignment of linked reads against a reference sequence all subsequent steps are based on the alignment file (BAM file) and that makes this the most important step of the pipeline. Any issue in initial read mapping will have consequences for the reliability of identified and reconstructed haplotypes. Therefore, it became clear that a non-haplotype aware reference assembly of a closely related accession should be used for alignment.

Planned results 2020	The overall aim of the project is to develop and perform haplotype phasing dedicated for polyploid genomes. As adaptations of existing software as well as development of a new pipeline appeared to be more difficult than foreseen the project timelines and deliverables had to be adapted. In addition there has been some delay of the project due to sample delivery issues. As a consequence we foresee that we still can finalize and reach most of the deliverables with the project ending 31st March 2020 instead of 31st December 2019. Still to be done and possible deliverables within project: * Target enrichment for population, all library preps and sequencing - one final attempt with optimized Cas9 protocol * Haplotyping enrichment regions varieties and population - one go with current version of the SDhaP haplotyping pipeline * Full genome haplotyping carrot - one go with current version of SDhaP haplotyping pipeline * Full genome haplotyping innovator - one go with current version of SDhaP haplotyping pipeline

	A proposal for project extension for further validation of the haplotyping results has been discussed with all participants but was not sufficiently supported by all companies to be granted.

Deliverables/products in 2019 (provide the titles and /or a brief description of the products/deliverables or a link to a website.
Scientific articles: None
External reports: None
Articles in professional journals/magazines: None
(Poster) presentations at workshops, seminars, or symposia. None
TV/ radio / social media / newspaper: None
Remaining deliverables (techniques, devices, methods, etc.): See planned results 2020