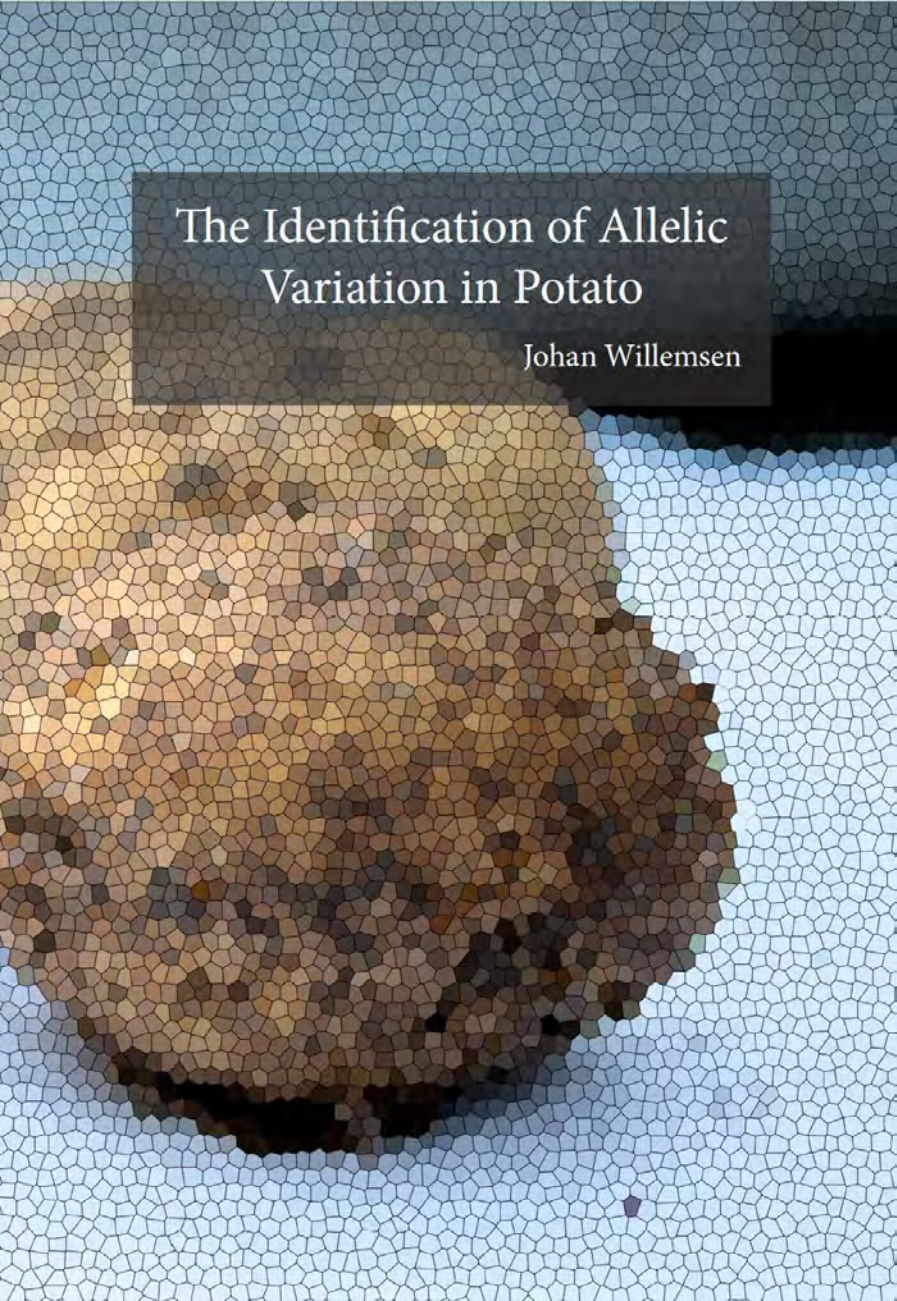
The Identification of Allelic Variation in Potato

Johan Willemsen



The Identification of Allelic Variation in Potato

Thesis committee

Promotor
Prof. Dr R.G.F. Visser
Professor of Plant Breeding
Wageningen University & Research

Co-promotor
Dr H.J. van Eck
Assistant professor, Plant Breeding
Wageningen University & Research

Other members
Dr J.M. de Boer, AVERIS Seeds B.V., Valthermond, the Netherlands
Prof. Dr F.A. van Eeuwijk, Wageningen University & Research
Prof. Dr D. de Ridder, Wageningen University & Research
Prof. Dr B.J. Zwaan, Wageningen University & Research

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences (EPS).

The Identification of Allelic Variation in Potato

Johan Willemsen

Thesis

submitted in fulfilment of the requirements for degree of doctor
at Wageningen University
by the authority of the Rector Magnifucs
Prof Dr A.P.J. Mol,
in presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 21 November 2018
at 11 a.m. in the Aula.

Johan Willemsen
The identification of allelic variation in the autotetraploid potato
207 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)
With references, with summary in English

ISBN: 978-94-6343-513-0
DOI: https://doi.org/10.18174/459655

Table of contents

Chapter 1

General Introduction

The potato (*Solanum tuberosum*) can be seen as a model system for both autopolyploid crops and as model system for the development of storage organs. Potato is a highly heterozygous autotetraploid species (2n = 4x = 48), which implies that each potato variety contains four distinct homologous chromosomes. Nowadays, most of the commercially available varieties are tetraploids. From a plant breeding perspective, the goal of a potato breeding programme is the development of improved varieties, which contain better or novel traits in comparison to existing varieties. In that regard, potato breeding is aimed at developing varieties that are high-yielding, but also contain disease resistances, and have excellent quality criteria. Genetic improvement of potato varieties traditionally has been achieved with phenotypic selection, as the genetic basis of many traits was poorly understood (Hamilton, 2011). Intriguing examples of these traits are plant maturity (e.g. daylight-dependent tuberization) and potato tuber shape (e.g. long or round tuber shape), where continuous trait variation is seemingly explained by multiple alleles at a single locus. To obtain better performing varieties an improvement of the genetic basis of these varieties is needed.

A typical potato breeding program consists out of crosses between highly heterozygous parents. After selection in the F1 generation, breeders hope to identify a few progeny outperforming their parents which can be released as variety or selected for further breeding. This whole process of repeated crossing and selection takes around 10 years. In the context of marker-assisted breeding, *molecular markers* can be used to characterise DNA variation that is predictive for important agronomical traits, with the sole purpose to speed up this breeding process. The identification of this molecular variation linked to trait variation is the foundation of marker-assisted breeding (MAS). These molecular markers are used to select parents with favourable allele composition, or select progeny that score positively for the presence of these markers.

Genetic studies in potato

Genetic studies in potato are hindered by the occurrence of polyploidy, coupled with self-incompatibility, and inbreeding depression. Although challenging, the availability of a reference genome (PGSC, 2011), and high-quality linkage maps (van Os et al. 2006; Hackett et al. 2013; Massa et al. 2015), allows to perform these genetic studies on a more routine basis. Among polyploid species generally a distinction is made between autopolyploids and allopolyploids. Potato is considered a autotetraploid species, as

tetraploidy originated due to whole-genome duplication in the distant past. Like most autopolyploid crops, potato displays polysomic inheritance, where any of the homologous chromosomes can pair with each other (Bourke, PhD thesis). In addition, potato exhibits severe inbreeding depression, which implies that breeders maintain a high degree of heterozygosity to obtain high-performing potato varieties.

Genetic mapping in experimental populations

In potato, for a long time genetic analysis focused on traits with simple genetic inheritance, due to complexities related to linkage mapping at higher ploidy level. Indeed, many of the early studies that performed genetic mapping in tetraploid F1 populations mainly identified QTLs for resistance, such as the H3 locus (Bryan et al. 2002), R2 locus (Li et al. 1998) and Ry_{sto} (Brigneti et al. 1997). These traits often display qualitative resistance, where a single dominant allele is needed for complete resistance. Recent progress in the development of tools for tetraploid linkage mapping (Hackett et al. 2002; Bourke et al. 2018) and the possibility to assign each marker to its putative homologous chromosome (Zheng et al. 2015), allows to model these QTLs not only by the contribution of each individual SNP, but rather on the relative contribution of each individual homolog. The application of these tools in tetraploid mapping panels allows to perform QTL mapping with more complex traits (Massa et al. 2015, Bourke et al. 2018). Simulation studies also suggest that power for QTL discovery might increase by using the presence or absence of each homolog as factor in QTL analysis, compared to single marker QTL analysis (Bourke et al. 2018). However one of the drawbacks of QTL mapping in such populations is lack of allelic diversity that can be screened within a single population (e.g. only eight homologs). Moreover, the mapping resolution depends on the frequency of recombination, and only with a large panel size a locus can be refined to a small region of interest.

Association mapping

In contrast to QTL mapping in F1 populations, a genome-wide association study (GWAS) measures the association between each marker and phenotype within a large panel of varieties. Such panels of varieties exhibit a high allele diversity, as they often are composed of a set of 'unrelated' varieties, displaying varying levels of IBD to each other. In principle, association mapping is analogous to QTL mapping in a F1 population, as in both cases the association between allele and phenotypes is measured. In the context

of association mapping, this association between marker and phenotype is based on the assumption of linkage disequilibrium of a marker with a QTL. The strength of the association between allele or marker and phenotype significantly depends on the proportion of phenotypic variance within the population explained by the marker, which in turn is a trade-off between frequency of such a molecular variant in the population and its allelic effect (Korte and Farlow, 2012). An very important aspect of association mapping is how reliable the individual marker-alleles tag the causative allele(s), as markers might be present in multiple haplotypes, diminishing the correlation between marker and phenotype.

The composition of the association panel does often influence the outcome of association mapping experiments. If traits are correlated with population structure, false-positive marker-trait associations might occur frequently (Kang et al. 2008; Rosyara et al. 2016). In potato population structure is exemplified in the identification of population structure in three separate population groups, processing, starch and 'rest' (D'hoop et al. 2011; Vos et al. 2015). A marker that is present at higher frequency in any of the structure groups could lead to a misleading marker-trait association, if phenotypic variation is also correlated with these structure groups. Therefore, the composition and selection of varieties for an association panel is crucial for the success of these studies. To avoid population structure, these panels can be selected by maximizing the genetic distance between accessions (Li et al. 2010). An alternative approach is to select a representative subset of varieties, balanced in every subpopulation, to minimize the effect of population structure. After performing association mapping, the identified QTLs need to be validated. Firstly, the QTL(s) can be validated in an independent association panel. Secondly, parents can be selected, in which this QTL is expected to segregate. Subsequent linkage mapping can be followed by a QTL mapping experiment to validate the effect of the QTL(s).

In potato several association mapping experiments have been performed, and this led to the identification of many QTLs related to agronomical traits (D'hoop et al. 2014; Schönhals et al. 2016; Sharma et al. 2018). Successful application of GWAS allowed to identify a single major effect QTL for plant maturity (Kloosterman et al. 2013). Likewise, application of GWAS to glycoalkaloid content allowed to detect multiple QTLs

associated with this trait (Vos et al. 2017). Many traits will show a polygenic genetic architecture, where each QTL will have a small effect on the trait.

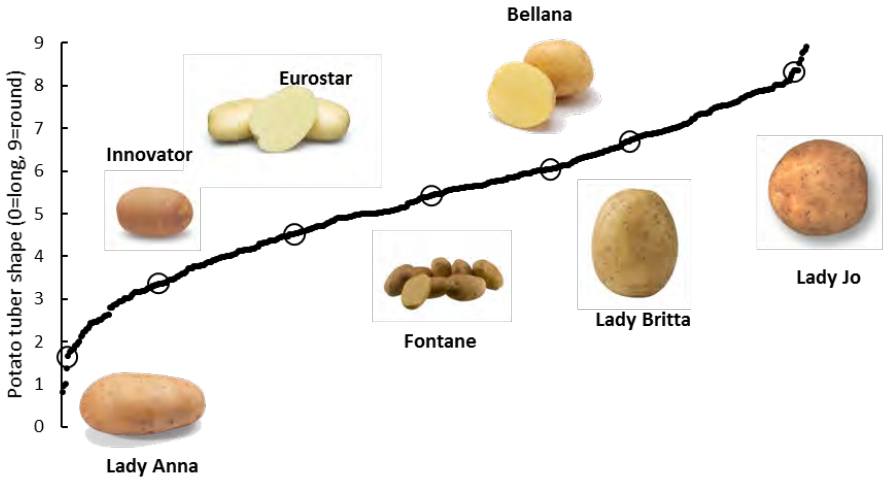


Figure 1. Phenotypic diversity for potato tuber shape in a panel of 537 potato varieties. Overall tuber shape in cultivated potato ranges from round to extreme long, but shows a continuous distribution.

Plant maturity and tuber shape

For potato plant maturity, initially a major effect QTL was found on chromosome 5, using a genome wide association study (GWAS). Subsequent linkage mapping in a diploid full-sib mapping population refined the locus. Gene function studies revealed that the *StCDF1* gene was responsible for regulation of day-light dependent tuberization (Kloosterman et al. 2013). Two dominant alleles of the *StCDF1* gene were found: Firstly, the *StCDF1.3* allele, containing a transposon insertion. Secondly, an excision allele, containing a 7bp footprint of the transposon insertion (*StCDF1.2*). Both these allelic variants result in a truncated *StCDF1* protein that evades post-translational light regulation, leading to early tuberization. Likewise, potato tuber shape is a quantitative trait, displaying tuber shapes ranging from round to elongated (See Figure 1). A substantial part of this phenotypic variation is mediated by a single major effect QTL on chromosome 10. For this QTL previously three alleles were identified, from which only one conferred an elongating effect (van Eck et al. 1994).

Multiple alleles

Various mechanisms can be envisaged that give rise to a quantitative genetic effect at a single genetic locus. The most simple explanation is that multiple loci jointly influence trait variation. An alternative explanation is that multiple alleles at a single genetic locus explain phenotypic differences. Hypothetically, in case of multiple alleles, allelic variants that are expressed at different level or have different enzymatic activities, can combine at a locus to produce subtle quantitative phenotypic differences. Furthermore, other complex combinations of expression and protein variation can be envisioned that could explain a fully quantitative genetic effect. Here a haplotype is defined as a segment of multiple adjacent SNPs that are present in only one homologous chromosome. Commonly an allele can be defined as a variant of a gene or locus.

Indeed, the central aim of this *PhD* thesis is to identify alleles that are associated with trait variation. The occurrence of multiple functional alleles for traits such as tuber shape, and plant maturity is the main reason to pursue the development of haplotype reconstruction methods, as single SNP markers do not allow to disentangle these alleles and their contribution to phenotypic variation. To investigate the effect of each haplotype and allele, or their interactions, a comprehensive overview of all allelic variation present at a locus is needed, which can be achieved by methods such as described in this thesis. Subsequently each of these discovered alleles, or combinations of alleles can be tested for association to a specific trait.

Molecular marker discovery

For genotyping of polyploid crops, several marker platforms are commonly used. Earlier studies in potato have used low-throughput marker systems such as SSR (Simple Sequence Repeat) markers or AFLP (Amplified Fragment Loci Polymorphism) markers (D' hoop et al. 2011). Recent studies have employed single nucleotide polymorphisms (SNPs) to characterise genetic variation (Vos et al. 2015; Fletcher et al. 2012). Generally, these SNPs are bi-allelic, allowing the distinction between two alleles. Each of these SNPs can have five different dosage classes in a tetraploid (AAAA, AAAB, AABB, ABBB and BBBB). These alleles are commonly coded using a binary number, where *0* refers to the reference allele, and *1* refers to the alternative allele (Figure 2). For a segment of n bi-allelic SNPs a total of 2^n haplotypes are possible. In a tetraploid these 2^n haplotypes can

be distributed in $\frac{(r+4-1)!}{r!(4-1)!}$ distinct combinations of four haplotypes, where r is the ploidy level, which in case of two bi-allelic markers results in 35 possible phasing configurations.

A

Haplotype	Binary coding
AA	00
AB	01
BA	10
BB	11

B

		2nd bi-allelic marker: AAAA	AAAB	AABB	ABBB	BBBB
		0	1	2	3	4
1 bi-allelic marker: AAAA	0	00	01	02	03	04
AAAB	1	10	*11*	*12*	*13*	14
AABB	2	20	*21*	*22*	*23*	24
ABBB	3	30	*31*	*32*	*33*	44
BBBB	4	40	41	42	43	44

Figure 2. A) Possible haplotypes given two bi-allelic SNP markers. B) Possible genotypic combinations of two bi-allelic SNP markers. For the genotypes depicted in grey, Haplotype inference is needed. For instance genotype 11 can originate from haplotype configuration 11|00|00|00 or haplotype configuration 01|10|00|00. The haplotype contribution of the other genotypes can be inferred directly (i.e. genotype 00: 00|00|00|00 or genotype 01: 01|00|00|00).

These genotyping assays rely on fluorescent signals which allow determination of these different dosage classes (Voorrips and Maliepaard 2008), for instance in high-throughput genotyping arrays (Vos et al 2015; Fletcher et al. 2012) or KASP markers.

Sequencing technologies

Sequencing-based genotyping has rapidly become a more important tool to characterise genomic variation, and perform SNP discovery. Several studies show that in potato it is feasible to obtain usable genotypic information using high-throughput short-read sequencing with Illumina (Uitdewilligen et al. 2013; Slater et al. 2014). A key requirement before downstream application, is the estimation of dosage classes with this read information. The estimation of a dosage of a SNP in an individual is based on relative counts of reads that belong to either the reference allele or alternative allele of that SNP. Commonly, a polyploid variant caller such as FreeBayes is used to determine these dosages (Erikson et al. 2008). Generally, these tools require a high read depth to reliably assign dosage classes. Previously it was determined that in practice a read depth of > 80× is needed to achieve a 95% accuracy (Uitdewilligen et al. 2013), mainly to distinguish between simplex (ABBB), duplex (AABB) and triplex (AABB) genotype calls. Currently,

short-read Illumina sequencing allows a read length up to 250bp. These read lengths are still limiting applications for haplotype detection or identification of structural variants.

In recent years, however, third generation sequencing technologies have become available allowing sequencing of very long DNA fragments. These long-read, single-molecule technologies such as those offered by Oxford Nanopore and Pacific Biosciences, potentially allow to sequence DNA molecules with lengths of several 100 kb. Application of these new technologies will allow to characterize highly complex regions, which are characterized by complex repeats and/or inversions. Examples of these regions are resistance gene clusters in plants (Witek et al. 2010), where complex repeat structures hinder interpretation and development of reliable marker assays.

Haplotypes are more suitable as markers than bi-allelic SNP markers

Nevertheless, most of these molecular marker technologies result in bi-allelic markers, which means that a single marker distinguishes only between two haplotypes. In many cases, multiple haplotypes are present at a locus, and a one bi-allelic SNP marker is not specific for a single haplotype. Obviously in an ideal situation, every marker has a high haplotype-specificity. One way to achieve haplotype-specificity is to consider multiple bi-allelic markers and reconstruct haplotypes. The reconstruction of haplotypes allows to improve distinction between different alleles. The application of these haplotype markers in QTL detection will likely result in a stronger association signal and therefore improve QTL detection power.

Allele mining

For diploids many tools have been developed that allow to characterise allelic variation, either by statistical phasing or by haplotype assembly (Browning and Browning, 2011), however efforts for phasing or assembly in polyploids have been limited, and studies that performed large-scale haplotype reconstruction have so far not been reported. Statistical phasing employs techniques to obtain haplotypes by exploiting information over multiple individuals. Conventionally these methods use unphased genotypes (i.e. dosages of individual SNPs), and estimate the most likely phasing for each individual, by exploiting information of the allele frequency of each allele, or identify-by-descent information. With the use of this information the most probable phasing is selected. Often in case multiple phasings are equally likely, one of the likeliest solutions can be

selected. Despite efforts in polyploid phasing no reliable and scalable method has been developed that allow phasing over a large number of variants. Existing approaches such as Satlotyper (Neigenfind et al. 2008), polyHap (Su et al. 2008) and SHEsis (Shen et al. 2016) only allow to phase a limited number of SNPs at high computational cost.

In contrast, haplotype assembly makes use of SNP-alleles that are jointly present in a single sequencing fragment, to allow phase estimation based on physical linkage in sequencing data (Figure 3). With increasing ploidy levels the complexity of estimating haplotypes increases, limiting the application of these methods in higher polyploids. Nonetheless, for polyploids several tools were introduced that allow to reconstruct haplotypes from short-read sequencing data (Anguir et al. 2013; Berger et al. 2014; Das et al. 2014; Xie et al. 2015). Based on a simulation study these approaches were found to have a high error rate, and often reconstruct chimeric haplotypes (Motazedi et al. 2016). In general these methods employ a reference-guided haplotype assembly and are performed within a single individual. In short, before haplotype assembly, sequencing fragments are aligned to the reference genome, after which genotyping is performed. Subsequently each alignment is interrogated for the presence of each SNP-allele, resulting in knowledge of physical linkage between SNPs (Figure 3). This information is exploited for the reconstruction of haplotypes. The advantage of sequencing-based haplotype reconstruction is the use of data of a single variety, avoiding the need for a large cohort of samples.

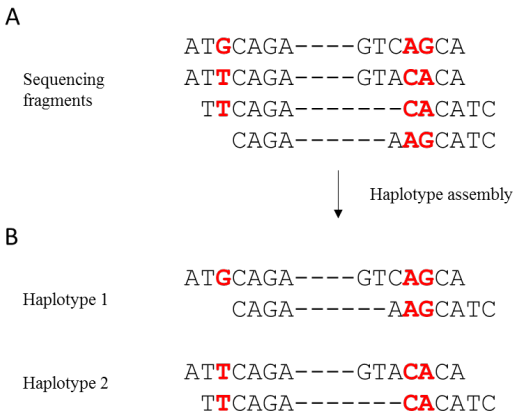


Figure 3. Haplotype assembly A) First sequencing reads are aligned. In red genomic variants are depicted. B) Based on the physical linkage of SNP-alleles within these reads two haplotypes can be defined. In polyploids also the copy number (dosage) of the haplotypes need to be determined.

Towards a haplotype map of tetraploid potato

In 2011 the potato genome was released, which was based on sequencing efforts in a doubled monoploid potato genotype (DM), where the final assembly comprised 722 Mb out of an estimated 831 Mb large genome (PGSC, 2011). As already mentioned above, the cultivated potato shows not only tremendous phenotypic diversity, but also has a highly heterozygous genome. Indeed, a single tetraploid variety has a SNP density of one variant in every 42 bp (Uitdewilligen et al. 2012), and exhibits extensive copy-number variations (Lovene et al. 2013, Hardigan et al. 2016). Despite targeted resequencing efforts to uncover allelic variation associated with agronomical traits (Schönhals et al. 2016; Uitdewilligen et al. 2012; Schreiber et al. 2015), knowledge about the number of alleles per locus and haplotype diversity in the tetraploid potato gene pool is limited. Efforts to start building a haplotype map of potato were started with the study of Uitdewilligen et al. (2012), where 129,156 genomic variants were identified in approximately 800 genes. The aim of that study was to a discover SNP markers for use on a SNP array. In addition, the sequencing data in this study was generated to explore the construction of haplotypes from sequencing data in tetraploid potato. The developed SOL-STW 20K SNP array was subsequently used for genotyping of 537 potato cultivars (Vos et al. 2015).

As potato has a highly heterozygous genome, we expect a high number of unique alleles in the potato gene pool. A single potato variety may contain an average of 3.1 unique alleles at any given locus, and because recombination leads to an exponential decay of linkage between alleles, a set of more frequent haplotypes (common), and an excess of low frequent (rare) haplotypes is observed. Based on the estimation of linkage disequilibrium between markers, and therefore linkage decay, potato should contain anywhere between 6-12 founder haplotypes (Vos et al. 2017). This suggests that the potato haplotype landscape is characterised by a mosaic of large haplotype blocks.

Outline of the thesis

In this thesis methods are described that allow the identification of haplotypes in panel of 'unrelated' varieties of polyploid crops. These haplotypes were subsequently used to shed light on the question, if multiple alleles contribute to trait variation. In addition, a step is taken towards the construction of a haplotype map of potato. This thesis addresses the weaknesses and challenges of methods for obtaining haplotypes, and covers their

application in polyploid genetic studies. The developed haplotype reconstruction methods allow to routinely characterize haplotype diversity in tetraploid potato.

In **Chapter 2** the inheritance of potato tuber shape and eye depth was studied, using a genome-wide association mapping that allowed to define a region of 3.1 Mb on potato chromosome 10 where a major effect QTL for overall tuber shape is located at similar genomic location as a major effect QTL for tuber eye depth. A recombinant screening for tuber shape within a diploid bi-parental mapping population (C×E), refined this region to a cluster comprising 277 kb of sequence. In this chapter only single marker analysis were used, limiting our ability to investigate haplotype structure for this region.

In **Chapter 3** a new method is described which reconstructs haplotypes from sequencing data. For this we used exome sequencing data of ~800 genes in 83 tetraploid potato varieties. To build these haplotypes a stepwise approach was used, where first short-range haplotypes are reconstructed, followed by a haplotype extension step that allows to compute longer haplotypes. The accuracy of this method was verified using simulated sequencing data. The reconstructed haplotypes allowed to determine haplotype diversity within potato.

In **Chapter 4** a new approach was introduced to reconstruct haplotypes from conventional SNP array data. This approach can be seen as complementary to the approach in chapter 3, as conventional sequencing data does not allow to estimate haplotypes over longer distances, due to limitations attributed to sequencing technologies. Here we perform long-range phase estimation by exploiting relations within a potato association panel. The method that is presented here consists of two steps: 1) Pairwise SNP phase estimation using the EM algorithm. 2) Constructing full-length haplotypes SNP by SNP. The accuracy of this tool was validated with experimentally obtained haplotype information of two amplicons of the *StGWD1* gene.

In **Chapter 5** the reconstructed haplotypes were used to explore haplotype-based association mapping. Haplotype-based regression was performed for four potato traits: plant maturity, flesh colour, tuber shape, and potato tuber uniformity. This led to the identification of not only SNP markers associated with these phenotypes, but also point towards which haplotype is responsible for differences in trait. From these results, a strong relation between the power of detecting a QTL and haplotype-specificity of a SNP

for a causative allele was observed. Using haplotype-based regression we could identify novel QTLs for plant maturity and flesh colour.

In **Chapter 6** a haplotype imputation method is presented that makes use of a haplotype library. This haplotype library is composed out of curated haplotypes, and imputation is aimed at identifying which unphased genotypes contain these haplotypes. This imputation method allows to improve the haplotypes obtained in both Chapter 3 and Chapter 4.

In **Chapter 7** the developed methods for haplotype reconstruction and imputation were applied to characterise the allelic diversity of the *StCDF1* gene in a set of 83 cultivars?

In the concluding **Chapter 8** the results of previous chapters are evaluated and put into a broader scientific perspective. The implications of these results for genetic research of potato are discussed.

Chapter 2

The *Ro* locus involved in potato tuber shape is located in a ~280kb region enriched for peroxidases

Johan H. Willemsen, Peter G. Vos, Arnoud Witteveen, Richard G. F. Visser, Herman J. van Eck

Abstract

Tuber shape is an intriguing morphological trait, which displays continuous trait variation ranging from flat, round to oval and long. Initially a single locus model was proposed to explain tuber shape, where multiple alleles rather than multiple loci were proposed to explain quantitative variation (van Eck et al. 1994). Besides this major-effect QTL on chromosome 10 another minor effect QTL has been published on chromosome 2, explaining 8% of the variance (Prashar et al. 2014).

To obtain a better overview on the loci contributing to variation in tuber shape a comprehensive genome wide association study (GWAS) was performed in a panel of 537 commercial potato cultivars. This confirmed that the *Ro* locus is the major-effect QTL, but also the minor effect QTL on chromosome *2* was found. In addition, on chromosome *10*, colocalization of the major effect QTL for tuber shape and a major QTL for eye depth was observed. For the *Ro* locus, most significant associations were found to localize on superscaffold PGSC0003DMB000000385 on chromosome 10.

To refine this region we performed a recombinant screening in a diploid population (C×E). Recombinant analysis resulted in the identification of 104 recombinants originating from the female meiosis and 27 recombinants from the male meiosis. Recombinant analysis with additional SNPs within the selected region allowed us to confine the *Ro* locus to a 280 kb region, located on superscaffold DMB546 (323kb). Within this region a cluster of cell wall III peroxidase genes is found. Based on the putative role of peroxidases, this gene family cluster of repeats is likely to be implicated in mediating differences in tuber shape.

Keywords

Solanum tuberosum, organ morphology, high resolution mapping, GWAS, haplotypes

Introduction

Cultivated plant species often display tremendous morphological variation in organ shape due to domestication and selective breeding (Alonso-Blanco et al. 2009). Well-studied examples of phenotypic plasticity are e.g. the variation in wheat grain size and shape (Gegas et al. 2010), tomato fruit shape (Monforte et al. 2014), or morphotype diversification in *Brassica rapa* and *Brassica oleracea* cultivar Groups (Cheng et al. 2016).

Likewise, the tubers of potato (*Solanum tuberosum* L.) display intriguing phenotypic variability for morphological traits such as eye depth, tuber size and tuber shape (Van Eck 2007; Li et al. 2005; Prashar et al. 2014). It is often assumed that such morphological traits are controlled by multiple genetic factors, because trait variation shows a continuous quantitative distribution, in line with the infinitesimal model of (Fisher, 1918) postulating that trait variation is controlled by multiple minor-effect loci.

In contrast, Sirks (1929) proposed a model where phenotypic variation is explained by multiple alleles of a single major-effect locus. Evidence for multiple alleles with gradations of effects has been provided for e.g. pinewood density (Groover et al. 1994), frost tolerance in Eucalyptus (Byrne et al. 1997), and flowering time in Arabidopsis (Johanson et al. 2000). In potato short-day-dependent tuber formation and tuber shape are controlled by a single major-effect locus with multiple alleles (Kloosterman et al 2013; Van Eck et al. 1994).

In this study we focus on potato tuber shape, which is a composite trait consisting of several aspects such as length/width ratio, curvature, eye depth, tapering and bulging (Van Eck 2007). In comparison to Latin American landraces, commercial varieties are highly uniform, only differing in tuber length/width ratio due to selection. This length/width ratio ranges from compressed (<1.0) to long (>2.0). From a market perspective the consumers prefer oval tubers, whereas processing industry prefers long tubers for the production of French fries, or round tubers for crisps.

Genetic studies of potato tuber shape, initially hindered by continuous trait values in tetraploid material, advanced with the use of diploids. A monogenic *Ro* locus was proposed, where round is dominant over long (Masson, 1985; Jong and Burns, 1993). Subsequently, the *Ro* locus involved in tuber shape was mapped on potato chromosome *10* in the diploid C×E population, with multiple alleles explaining more than 80% of the

genetic variance (Van Eck et al. 1994). Since then, several studies confirmed this major-effect QTL on chromosome *10* (Li et al. 2005; Śliwka et al. 2008; Prashar et al. 2014; Lindqvist-Kreuze et al. 2015), but additional minor-effect QTLs have been reported as well. Śliwka et al. (2008) describe two minor-effect QTL on chromosome 2 and chromosome 11, explaining 8.0% and 5.6% of the phenotypic variance, respectively. Li et al. (2005) proposed a locus involved in eye depth (*eyd* locus) at 4 cM distance from the *Ro* locus. The association between eye depth and tuber shape on chromosome 10, and the minor-effect QTL on chromosome *2*, have also been described by Prashar et al. (2014).

In this paper we analyse potato tuber shape, defined as length/width ratio from a genetic perspective. We use the recently published SOL-STW 20K Infinium SNP array (Vos et al. 2015) on a comprehensive association panel of tetraploid varieties. In addition a high resolution map and marker saturation was obtained using the C×E bi-parental diploid mapping population. This paper presents the high resolution mapping of the *Ro* locus in the C × E mapping population along with a genome wide association study (GWAS) using commercial tetraploids, and pinpoints the *Ro* locus to a 280 kb cluster of peroxidase genes. The putative role of peroxidase genes in modulating tuber shape morphology is discussed.

Materials and methods

Genome wide association analysis of tuber shape

For the purpose of this publication tuber shape is defined by the length/width ratio and all other aspects are ignored. Phenotypic data from a comprehensive panel of 537 varieties (Table S1) were collected from 221 varieties and 190 advanced breeders clones grown on trial fields at different locations with clay and sandy soil in four-hill plots during the normal growing season in 2006 and 2008 (D'hoop et al. 2011). In addition, for 299 varieties multi-year, multi-location data were available from breeding programs (D'hoop et al. 2011). In all cases the visual observations were recorded using an ordinal scale (1=long, 3=long/oval, 5=oval, 7=round/oval, 9=round) or converted to this scale. From these data sources the Best Linear Unbiased Estimators (BLUEs) for tuber shape were calculated using GenStat, release 8.11 (VSN International Ltd., Oxford, UK) and are shown in Table S1. BLUEs for Eye Depth were available for 190 out of the 221

varieties (D'hoop et al. 2011) following an ordinal scale from 4=very deep to 8=very shallow.

A Genome Wide Association study (GWAS) was done with linear mixed models (LMM) using BLUEs for tuber shape as response variable and allele dosages per marker in a genotype as fixed effect. Thus, trait variation was modelled with a strict additive genetic model assuming a linear dose-response between phenotypes and allele dosage. In addition, a general model was fitted, testing for significant differences in the means of the trait values between genotype classes (dosages).

Greater certainty about the putative presence of minor-effect QTL affecting tuber shape, beyond the major-effect *Ro* locus on chromosome *10*, was obtained by co-factor analyses. Cofactors increased the statistical power for QTL detection and allow identification of marker-trait associations at positions elsewhere in the genome. For significant markers located in the *Ro* locus a cofactor analysis was done by including markers as additional fixed effect predictors to the LMM model.

In addition we employed a multi-locus stepwise regression to identify the minimum subset of markers explaining most phenotypic variation within the detected QTL region on chromosome 10. Initially all markers located within this region were used to calculate pairwise linkage disequilibrium (LD), after which from each set of highly correlated markers ($r^2 > 0.9$), one representative marker was chosen. To select the best model we employed backwards selection where the least significant marker in the model was removed until all markers were significant ($p < 0.05$).

Apart from naive association studies, we also included a correction for population structure. A kinship matrix (K) was calculated using ecological distance in GenStat (1-|xi-xj|/r, unless xi=xj=0, where x_i and x_j are allele dosages and r is the range) on a subset of 710 markers selected on the basis of independence (Vos et al. 2016)

All genome-wide association studies using Linear Mixed Models (LMM) were done using the 'efficient mixed-model association' EMMA (Kang et al. 2008) as implemented in the GWASpoly R package (Rosyara et al. 2016). All naive regression analyses were done using the LM procedure in R (R Development Core Team, 2008). The explained variance for each marker is calculated by using the squared correlation coefficient between marker and phenotype.

An experiment-wide type I error of 5% was obtained by a Bonferroni correction. With 14,530 markers this Bonferroni significance threshold is $p < 3.44e\text{-}06$, which corresponds to a $-^{10}\log P$ value of 5.46. In view of Linkage Disequilibrium (LD) patterns not all statistical tests are independent, leading to an overly conservative threshold (Johnson et al. 2010, Gao et al. 2010). Hence, we also employed a threshold of $-^{10}\log P$ value of 4 to allow detection of minor-effect QTLs.

Haplotype structure at the Ro locus

To investigate the haplotype structure around the *Ro* locus we calculated the Pearson's r^2 between SNP allele dosages at a conservative threshold of $r^2 > 0.2$. These pairwise estimates were visualised in a 3 Mb window surrounding the *Ro* locus. Boundaries for haplotype blocks were manually placed and all markers were clustered using UPGMA.

High resolution mapping of the Ro locus

For the high resolution mapping experiment true seeds were sown from a cross between USW 5337-3 × 77-2102-37, hereafter referred to as C × E (Jacobs et al. 1995). At seedling stage the crumpled segregants due to the *Cr*-locus (Jongedijk et al. 1990) were removed. The remaining vigorous seedlings were planted out in boxes following a 12 x 8 grid. Leaf tissue was sampled in deep-well microtiter plates for a 'quick and dirty' DNA extraction using the NaOH-Tris method (Collard et al. 2007). Recombinants were selected from 2500 seedlings grown in two batches of 1500 and 1000 seedlings each. The 173 recombinant seedlings from the first batch and 29 recombinant seedlings from the second batch were transplanted to 1.1L pots and grown until senescence to harvest tubers. Meanwhile, new leaf tissue of recombinants was re-sampled for high quality DNA extraction using the KingFisher® genomic DNA purification kit (Thermo Scientific, Breda, The Netherlands) according to the manufacturer's procedures. Tuber shape of each seedling was visually classified as compressed, round, and long. Furthermore, the length/width ratio (L/W) of three representative seedling tubers was recorded.

PCR markers for recombinant analysis in C x E offspring are documented in Table S2. Intron spanning primers were selected with the Primer3+ program (Untergasser et al. 2007) using exon sequences from the potato reference genome (PGSC, 2012) as template. PCR reactions started at 94ºC for 3 minutes, followed by 40 cycles of 30 sec at 94ºC, 30 sec at 60ºC and 1 min at 72 ºC, and a final extension stage at 72ºC for 5 min using high

fidelity Phire© DNA polymerase in the presence of LCGreen™. PCR products were analysed using the Lightscanner (Idaho Technology Inc., Salt Lake City, UT). followed by melting curve analysis (De Koeyer et al. 2009). All markers were tested on 20 progeny of C x E and two replicates of each parent.

Ideally, markers were selected which display four melting curves to allow complete classification of the segregating alleles to identify recombination events both in the female and male meiosis. The general Mendelian model of segregating marker alleles in this BC_1 population is $ab \times bc \rightarrow ab : ac : bb : bc$, where the melt curves of *ab* and *bc* offspring can be identified using the *ab* and *bc* parental samples. The curves with the highest melting temperature, due to the absence of SNPs, identifies the *bb* class. The remaining group of curves should thus be the *ac* class. In this way *n* offspring allows to scan *2n* meiosis for recombination events (Figure S1). All seedlings were tested with markers LS_B466 and 495_499, designed at a safe distance flanking the most significant marker-trait associations located in superscaffold DMB385. The other markers (Table S2), developed to saturate this genetic interval, were genetically mapped using the recombinant offspring.

Functional analysis of candidate region

A physical region of 280 Kb comprising 13 candidate genes with a PGSC annotation was analysed. Dot plot analysis of the corresponding sequence from the DM reference genome was performed with Gepard (Krumsiek et al. 2007) to allow identification of repeat structures. Regions with potentially unannotated peroxidase genes were identified using BLASTn. For each hit a gene prediction was done using FGENESH (Solovyev et al. 2006). Differential expression of candidate genes expressed as fragments per kilobase of exon model per million mapped reads (FPKM) in tuber and other tissues was studied using RNAseq data (Massa et al. 2011) and the Potato Genome Sequencing Consortium (2011) for DM1-3 516 R44 and RH89-039-16 (hereafter referred to as DM and RH).

Results

Distribution of tuber shape in varieties

BLUEs for tuber shape trait values (Table S1) of a representative panel of 537 varieties (Vos et al. 2015) displayed a normal distribution (Figure S2). Within this set of varieties

extreme variation in overall shape is observed clearly displaying a normal distributed phenotype. Eye depth displayed a skewed distribution due to selection against deep eyes. The distribution observed among varieties will differ from the distribution in seedling generations, as values below a threshold will be regarded as unmarketable. A highly significant correlation (r = -0.57, N=190, p < 0.001) was observed between tuber shape and eye depth (Figure 1), where a long shape is associated with shallow eyes.

Different market niches require varieties with a specific tuber shape, and earlier research has shown that market niche is strongly confounded with the genetic structure of the gene pool (D'hoop et al. 2008; 2010; 2014; Uitdewilligen et al. 2013; Vos et al. 2015, Rosyara et al. 2016). Therefore, the association between the structure groups 'Starch, Agria and Rest' (Vos et al. 2015) and tuber shape and eye depth was studied. Varieties used by starch industry are significantly more round (6.84 ± 0.14) than varieties from the other structure groups Agria (5.01 ± 0.18) and Rest (4.98 ± 0.08) (Table S3). Eye depth is also significantly confounded with structure groups, where starch varieties are hardly selected for eye depth and deep eyes are common (5.08 ± 0.16), processing varieties have shallow eyes (6.76 ± 0.074), whereas the rest group is more diverse (6.20 ± 0.067) (Table S4).

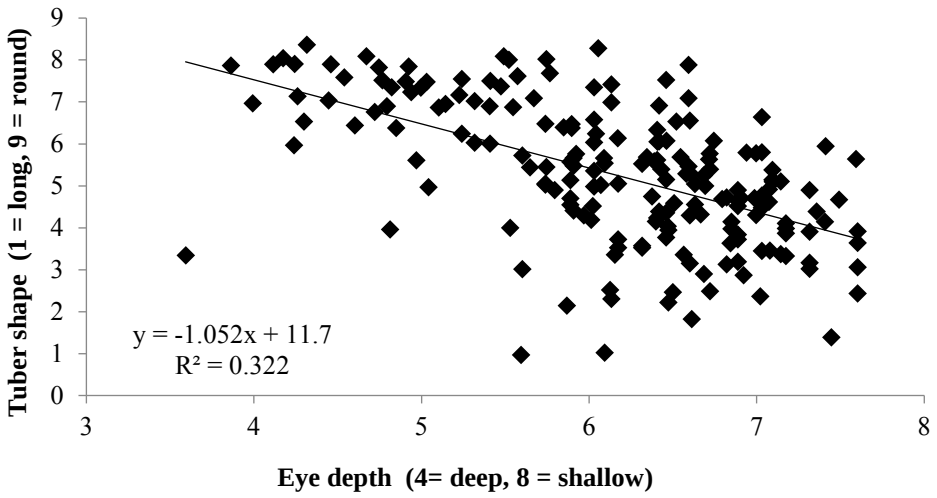


Figure 1. In a panel of distantly related varieties a significant correlation is observed between tuber shape and eye depth

Genome Wide Association Study

To identify QTL involved in tuber shape we applied GWAS to the phenotypic values (BLUEs) and 14,530 SNP genotypes (SolSTW SNP array; Vos et al. 2015). Initially a naive association model was fitted, ignoring the effects of population structure. This resulted in the identification of 346 SNPs which exceeded the Bonferroni corrected significance threshold ($-^{10}\log P \geq 5.46$). The corresponding Manhattan plot (Figure S4) suggests a single highly significant locus on chromosome *10* along with other significant associations scattered across the genome.

After inspection of diagnostic Q-Q plots (Figure 3) severe *P*-value inflation was observed, indicating spurious associations for a large number of markers. To reduce the number of false-positive associations a population-structure corrected LMM association was performed (Figure 2A). This resulted in a strong decrease from 346 to eleven significant marker trait associations, confirming the aforementioned correlation between structure groups and tuber shape. Also the *P*-value inflation was no longer problematic as judged from the Q-Q plot (Figure 3). Therefore, population-structure corrected models for association mapping are used hereafter.

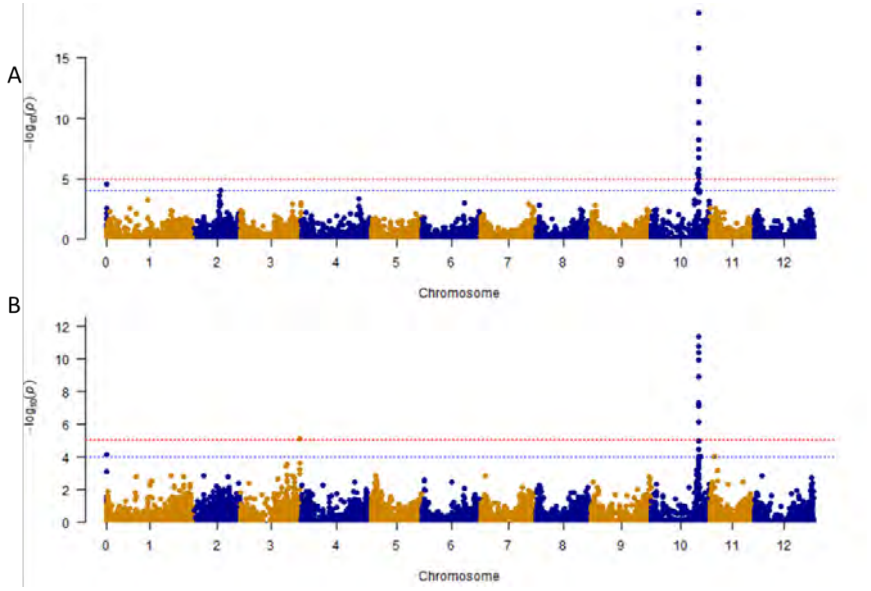


Figure 2. Manhattan plots for tuber shape and eye depth**.** Manhattan plots of a (A) kinship corrected GWAS of tuber shape with 537 genotypes, (B) kinship corrected GWAS of eye depth in 537 genotypes. Dotted (red) horizontal line is at the Bonferroni multiple-testing threshold of –10log(p) of 5.46. The blue dotted line is at the threshold of –10log(p) 4.0.

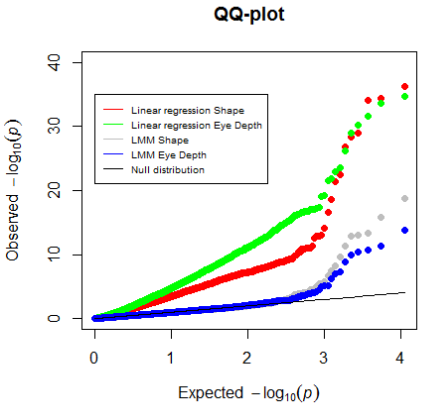


Figure 3. QQ-plots for association analysis.

These eleven SNPs (Table 1), exceeding the Bonferroni significance threshold (-10logP ≥ 5.46), indicate a single major-effect QTL on chromosome 10, exactly at the position where the *Ro* locus was expected. When a less stringent threshold was used ($-^{10}\log P \geq 4$) another nine SNPs were identified at the *Ro* locus, except for solcap_snp_c2_56344. This SNP is located at coordinate chr00:22129989, and tagged the unanchored 134 kb scaffold PGSC0003DMB000000773. On chromosome *2*, just below the significance threshold, two SNP markers were identified, PotVar0123847 at 27.60Mb ($-^{10}\log P$ =3.58) and solcap_snp_c1_11556 at 28.04Mb ($-^{10}\log P$ = 3.99), suggesting a putative minor-effect QTL. The SNPs solcap_snp_c1_5091 at 28.83Mb and solcap_snp_c2_51115 at 29.67Mb, reported by Prashar et al. (2014) were not associated at all with tuber shape in our material.

When the outcome of the 'general' model was compared with the 'additive' model we did not identify other significant SNPs, although $-^{10}\log P$ values varied slightly between these models. The putative minor-effect QTL on chromosome *2* could not be identified with the general model.

Marker	Chrom	Position	LMM: Shape additive (-10logP)	LMM: Eye depth additive (-^{10}logP)	Explained variation tuber shape (r^2)	Explained variation Eye depth (r^2)	MAF
PotVar0111687	10	48721966	18.71	13.83	0.26	0.29	0.29
solcap_snp_c2_25471	10	48808404	15.81	11.36	0.25	0.29	0.28
solcap_snp_c2_25485	10	48737840	13.34	9.91	0.25	0.26	0.28
solcap_snp_c2_25522	10	48617457	13.02	10.37	0.21	0.25	0.29
solcap_snp_c1_8019	10	48863165	12.81	10.74	0.21	0.27	0.27
solcap_snp_c2_25526	10	48617149	11.34	8.90	0.20	0.23	0.28
solcap_snp_c1_8020	10	48863048	9.63	3.72	0.17	0.13	0.24
solcap_snp_c1_16351	10	48761642	8.23	2.94	0.16	0.12	0.24
solcap_snp_c2_25532	10	48591792	7.40	3.98	0.14	0.11	0.24
solcap_snp_c2_25549	10	48875383	6.73	7.06	0.11	0.21	0.04
solcap_snp_c2_45611	10	48203431	5.76	3.67	0.13	0.10	0.24
PotVar0132241	10	47156274	5.38	1.55	0.08	0.01	0.02
solcap_snp_c1_8021	10	48862950	5.38	2.61	0.08	0.08	0.21
solcap_snp_c1_8018	10	48863220	5.05	7.27	0.08	0.17	0.03
solcap_snp_c2_25529	10	48593621	4.65	3.86	0.10	0.10	0.23
solcap_snp_c2_56344	0	0	4.58	3.10	0.07	0.08	0.19
PotVar0132243	10	47156328	4.49	0.97	0.08	0.01	0.02
solcap_snp_c2_55861	10	46117955	4.11	3.18	0.10	0.13	0.25
solcap_snp_c2_25527	10	48616993	4.03	2.62	0.08	0.09	0.23
solcap_snp_c1_11535	10	49553136	4.01	2.14	0.07	0.07	0.20
solcap_snp_c1_11556	2	28040094	3.98	1.77	0.01	0.01	0.08

Table 1. Significance and explained variance (R^2) of markers associated with tuber shape and eye depth (using LMM, additive model)

To gain more insight in putative minor-effect QTL we used cofactor analyses (Kang et al. 2008; Segura et al. 2012), to correct for the major effect of the chromosome *10* locus on associations elsewhere in the genome. For all significant markers on chromosome *10* cofactor analysis were performed, and in line with expectations the previously observed minor-effect QTL on chromosome *2* showed an increased significant association , while simultaneously the significance of marker-trait associations at the *Ro* locus was diminished.

Surprisingly, cofactor analysis using the chromosome 10 markers, did not cancel the effect of the chromosome 10 locus. Even using a cofactor on the most significant SNP (PotVar0111787) did not completely cancel out the effect of the *Ro locus*, but identified an additional marker-trait association with solcap_snp_c2_57634, located 0.6 Mb distal of PotVar0111787, with -^{10}log$P$ of 4.9 (previously -^{10}log$P$ of *3.88*), having allele frequency of 2% (Figure S5). All other cofactor analyses only partially cancelled the chromosome 10 QTL, whereas other markers surrounding the *Ro* locus either show a decreased or increased p-value, suggesting the occurrence of allelic heterogeneity, or a causal haplotype that is only partially represented by all SNPs (or multiple SNPs).

To investigate how much of the phenotypic variation was explained by SNPs located in the major QTL region, a multi-locus stepwise regression with backwards selection was

performed on markers located in the neighbourhood of the major QTL on chromosome 10. A total of 168 markers were selected, and pairwise correlations between these markers were calculated. From each set of highly correlated markers ($r^2 > 0.9$) one representative marker was selected, resulting in a total of 127 independent markers. Using this approach a total of 41% of the phenotypic variance was explained, by a total of 17 SNPs within the *Ro* locus. In addition we performed the multi-locus regression on the top 20 significant markers ($-^{10}\log P > 4$), resulting in 7 markers, explaining a total of 36.7% of the variation (Table S5). In addition, a weak putative minor QTL was identified on chromosome *4 (*$-^{10}\log P$3.3)*,* for which associations were found in the single SNP LMM, although only one SNP was identified with $-^{10}\log P$ higher than 3.7.

Based on these results we conclude that in this comprehensive variety panel a small number of SNPs show a highly significant association with tuber shape. These SNPs localize in a 2.4 Mb region delimited by PotVar0132241 and solcap_snp_c1_11535 at PGSC coordinates chr10:694962745 to 697359607, indicating the physical position of the major-effect *Ro* locus. In addition of this locus, a minor effect QTL was found in proximity of the QTL location reported by Prashar et al. (2014).

Haplotype blocks at the Ro locus on chromosome 10.

For each marker on chromosome 10 we calculated the Pearson's r^2 between SNP allele dosages. First, we explored the haplotype structure flanking the most significant marker on chromosome 10 (PotVar0111687). This revealed one haplotype block in which all observed marker trait associations are present. This block represents a physical distance of 2.3 Mb from PGSC0003DMB000000673 to PGSC0003DMB000000446 (Figure 4B, Figure 5) Subsequently we looked at pairwise distances of markers significantly associated with tuber shape, disentangling these 20 markers to a total of 5 distinct haploblocks.

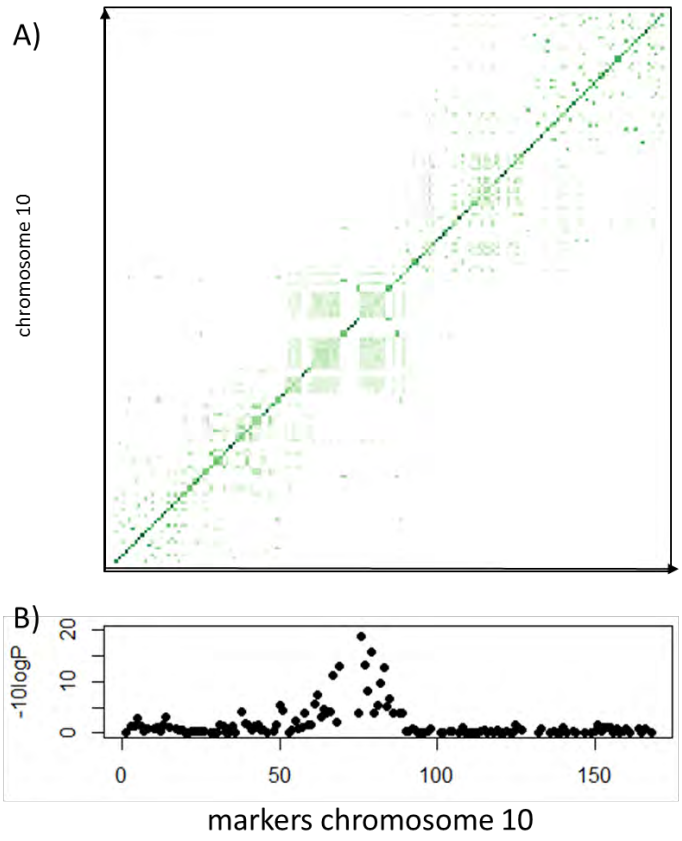


Figure 4. Haploblock structure of a 2.3 Mb region of chromosome *10* flanking the *Ro* locus. Linkage disequilibrium (LD) plot of markers located the chromosome 10 QTL. Manhattan plot of 168 markers suggesting that all significant markers are located in only one haploblock.

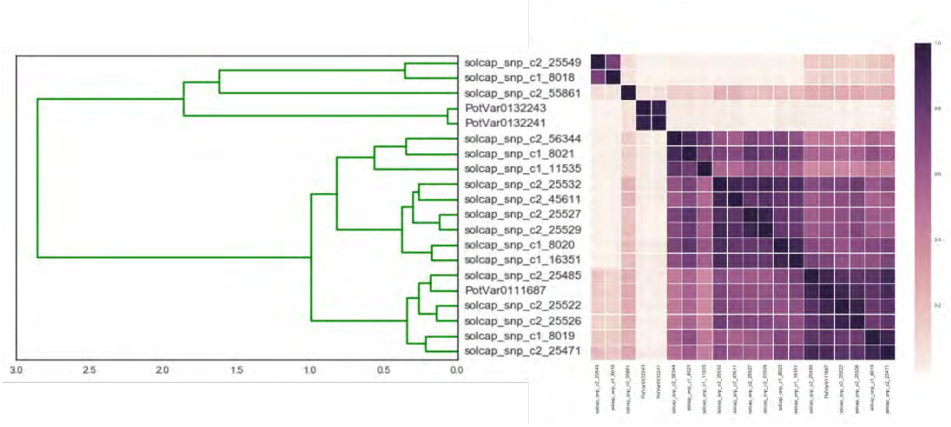


Figure 5. UPGMA plot of correlations between SNP markers significantly associated with tuber shape.

Integration of superscaffold DMB773 in the physical map comprising the Ro-locus

The most significant markers on chromosome 10 identified in the association analyses are located on superscaffold DMB385, but this provides little information on the boundaries the physical and genetic interval comprising the *Ro* locus, as between each superscaffold in the potato genome a (artificial) gap is inserted. In our GWAS a marker-trait association was observed on a unanchored scaffold DMB773 with marker solcap_snp_c2_56344 showing an association of $-^{10}\log P$ value of 4.58 using a kinship-corrected GWAS.

To explore the contiguity of the candidate region we compared the candidate region spanning DMB385 to DMB4446 using ENSEMBLCOMPARA (Vilella et al. 2009). to the tomato synteny. These observations suggested that unanchored scaffold DMB773 might be linked with scaffold DMB385 and belong to the candidate region for the *Ro* locus (Figure S9). In an effort to investigate the gaps and connections between PGSC superscaffolds, we used gap-spanning BAC end sequences from the RHPOTKEY and DM BAC library (Xu et al. 2011) = PGSC). A total of four BAC clones were found that connect DMB385 to DMB773, but none of these BAC clones connect DMB773 to DMB546 (Table S6).

In addition dot plot analysis suggests that repeat structures in and DMB385 and DMB773 overlap (nonspecific lipid transfer proteins). Flanking nonspecific lipid transfer proteins (DMG31237 and DMG 31236) at the end of DMB385 display 99% similarity to NSLTPs (DMG11951 and DMG11952) at the start of scaffold DMB773. Likewise two peroxidase genes were found in DMB773, which have high similarity to the peroxidase cluster on DMB546 (Table S7).

High resolution mapping and marker saturation

To fine map the region containing the *Ro* locus, we developed markers LS_B446 and 495_499, spanning an interval of 2.7 Mb on chromosome *10* which comprise the complete haplotype block observed above. These boundaries were chosen at ample distance from scaffold DMB385, where the most highly associated SNPs were identified, to ensure inclusion of the tuber shape locus. Genotypic data collected from 1472

seedlings with the marker loci LS_B446 and 495_499 allowed us to identify 142 maternal and 31 paternal recombinants (5.9 cM averaged).

Subsequently a fine mapping study was done in the CxE diploid population. Initially 1070 seedlings were grown and we identified 142 maternal and 31 paternal recombinants. After three months of growing, late maturing recombinants did not have tubers, hence the final number of informative meiosis, including double recombinants, was 105 and 27 respectively in the female and male parent. Tuber formation on these replanted recombinant seedlings allowed an initial visual classification of round and long descendants.

Another set of seven HRM markers in the 2.7 Mb interval between LS_B446 and 495_499 allowed to fine map the Ro locus and to count maternal and paternal recombination events in each interval as shown in Figure 3B. No recombination events were observed between tuber shape and markers Asp6678 and Per20801. Therefore the flanking markers PhoTr31222 and Amt241073 were used to screen the second batch of 1000 seedlings to improve the genetic resolution in the remaining 1.1 Mb interval between PhoTr31222 and Amt241073. The second batch provided 19 maternal and 10 paternal recombinants in 798 informative descendants. These additional recombinants allowed to map the Ro locus between markers Asp6678 and Per20801, which both reside on superscaffold DMB546.

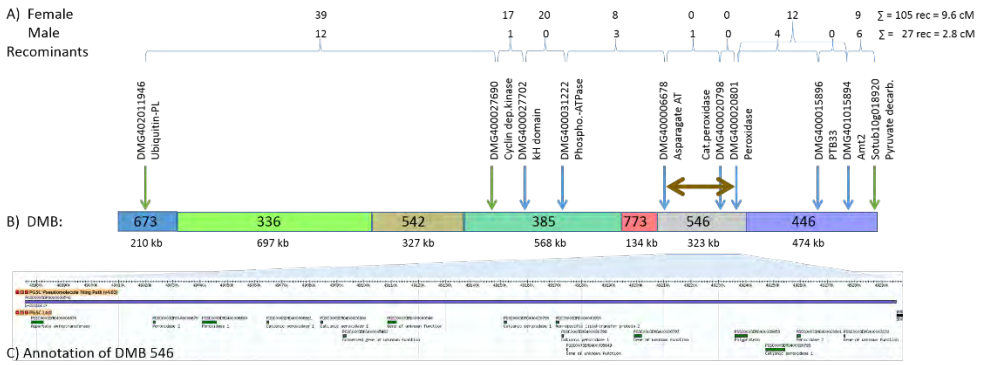


Figure 5. Overview of the region containing the *Ro* locus. A) Recombinants and B) Physical map of a 3 Mb region chr10:46970612-49818972 showing DMB scaffold numbers and positions of PCR markers for recombinant analysis. C) The majority of gene annotations found within the 280 kb region underlying the *Ro* locus belong to the large gene family of peroxidases.

Phenotypic analysis of tuber shape in the CxE

Within the CxE progeny two different shape categories were found; round and elongated. Visual assessment of harvested tubers allows classification of descendants into two groups, but compressed phenotypes (LW <1) can also be distinguished. After senescence the 173 recombinant seedlings were phenotyped, however due to effects of maturity only 128 genotypes did tuberize. The fact that all of these genotypes co-segregated with tuber shape allowed us to estimate the effects of different genotypic classes on tuber shape.

The batch of 128 co-segregating recombinant offspring segregated in 90 round and 36 long descendants, fitting a 3:1 Mendelian ratio (Chi-square= 0.86; p = *0.35, p > 0.05*) according to the model $Ro/ro \times Ro/ro \rightarrow Ro/\cdot : ro/ro$.

These 128 recombinant seedlings were also classified according to the melt curves of HRM markers into four groups AB : AC : BB : BC which descend from a cross between AB × BC parents. The comparison between phenotypes scored by length/width ratio to visual assessment, misclassification is possible between a length/with ratio of 1.2 and 1.5. The BB group with average L/W 1.08 coincides with offspring with long tubers. The compressed phenotypes typically belong to the AC group with average LW<1. Using the four genotypic classes ($Ro^{\varphi}Ro^{\sigma}$:$Ro^{\varphi}ro^{\sigma}$:$ro^{\varphi}Ro^{\sigma}$:$ro^{\varphi}ro^{\sigma}$) as identified from classification with HRM, a significant difference between the recessive long (ro/ro) and genotype classes having a dominant Ro allele (Ro/ro). An analysis of variance (ANOVA) indicated that Ro/Ro in general is distinguishable from categories with one Ro allele. Also no significant difference between the effects of both round alleles was found (Table 2).

Genotype	Mean	Tukey
$Ro^{\varphi}Ro^{\sigma}$	0.85	**a**
$ro^{\varphi}Ro^{\sigma}$	0.94	**ab**
$Ro^{\varphi}ro^{\sigma}$	1.08	**b**
$ro^{\varphi}ro^{\sigma}$	1.69	**c**

Table 2. Phenotypes per genotypic class based on batch CE2013-aprwith 128 recombinant seedlings

The recombination landscape of the genomic region near tuber shape

In the first batch of 1500 seedlings 105 maternal recombinants were observed between flanking markers. The physical distance corresponds to a region of 2.7 Mb, which relates

to a genetic distance of 10 cM. Relating this to the physical distance, on average one recombination event occurred every 25 kb. Less recombinants were identified in the male parent, where in this 2.7 Mb large region only 27 recombinants were identified, corresponding to a genetic distance of 2.5 cM. However in the 280 kb large region between markers Asp6678 and Per20801 no recombination events were observed. Recombinant analysis in the second batch was performed on a smaller interval of 1.1 Mb, resulting in distance of 2 cM in the female parent, and 1.3 cM in the male parent. The region of 280 kb corresponds to a genetic distance of 0.62 cM in the female parent and 0.12 cM in the male parent in the second batch. However in the first batch cosegregation was found for this region, resulting in an average genetic distance of 0.267 cM and 0.054 cM in respectively female and male parent. All information about fine mapping can be found in Table S8.

Candidate genes underlying the Ro locus

DNA sequence analysis of the candidate region of the tuber shape locus in DM identified two repeat clusters within scaffold DMB546, consisting out of number of cell wall type III peroxidases (Figure 5), three genes with unknown function (DMG35649, DMG20797, DMG45482), a non-specific lipid protein (DMG40954), a polyprotein (DMG39458), and an aspartate aminotransferase (DMG6678).

Gene expression data from Massa et al. (2011) and the Potato Genome Sequencing Consortium (2011) was checked for expression of these candidate genes in DM (elongated tubers) and RH (oval tubers). Based on this analysis the prediction of DMG35649, DMG45482 and DMG39458 is not supported by expression. A domain-search on polyprotein DMG39458 suggests that this gene is in fact an incorrectly annotated transposon.

Inspection of the dot-plot self-alignment of the DMB546 scaffold showed additional tandem arrays of unannotated sequences with homology to annotated peroxidase genes (Figure 5C). For these unannotated sequences a gene prediction was made using FGENESH (Salamov et al. 2000). Alignment of these new predictions with existing peroxidase genes from scaffold DMB546 supported the prediction of additional peroxidase genes. Information of genes and location of unannotated genes can be found in SI table 2,3.

Discussion

Tuber shape is an important morphological trait for potato breeders, relevant for utilization further down the product chain, but also intriguing from a fundamental perspective. We used an comprehensive association study in a large variety panel, coupled with fine mapping in a bi-parental population to narrow down the region in which the major QTL for tuber shape on chromosome 10 is found (van Eck et al. 1994). Despite all efforts no plausible minor QTLs were identified in our variety panel. The high resolution mapping significantly narrowed the genetic of the *Ro* locus to approx. 0.5 cM anchored within a 280kb region within DMB546 enriched for peroxidase genes.

The major QTL on chromosome 10 observed in our association panel is certainly at the same position as reported by Van Eck et al. (1994), Li et al. (2005), Prashar et al. (2014), Lindqvist-Kreuze et al. (2015). Using the most significant markers as cofactors does nullify the effect of the chromosome 10 QTL, but allows the detection of a minor QTL on chromosome 2. Intriguingly all marker trait associations in this regions are 'old' variation, introduced in the oldest variety (Yam), indicating that most functional variation for tuber shape is already present long time in the potato germplasm and is not due to recently introgressed haplotypes. The presence of minor alleles leads to an increase in roundness of the tuber shape.

Recently an association study was published by Rosyara et al. (2016) in 187 varieties, resulting in an association for tuber shape for the same chromosomal location at SNP solcap_c1_8019 using the solCAP diversity panel (Felcher et al. 2012), whereas tuber shape had a broad-sense heritability of 0.94. No additional minor QTLs were detected for shape, which might be explained by the small panel size (N=187) of this GWAS.

In addition the putative minor effect QTL involved in tuber shape on chromosome 2 at 28.04Mb maps at only 0.79 Mb distance from the QTL involved in tuber shape as identified by Prashar et al. (2014) and possibly Sliwka et al (2008). The most significant SNPs (solcap_snp_c1_5091 at 28.83Mb and solcap_snp_c2_51115 at 29.67Mb) as reported by Prashar et al. (2014) were not associated with tuber shape in our material, although based on co-localization, it is very likely that this QTL is the same position as previously reported. Prashar et al. 2014 speculates that introgressed late blight resistance originating from *S. demissum*, the elongating long allele on the chromosome two minor

QTL might originate from *S. demissum.* Nevertheless, the most significant marker within the chromosome 2 (solcap_snp_c1_11556) QTL does have a moderate MAF of 0.08, and in our panel this SNP is present in both old heirloom, as modern varieties (i.e. old variation) (Vos et al. 2016).

Relation with Eye Depth

We report the correlation between eye depth and tuber shape, both co-localizing on chromosome 10. Previously this correlation was observed before in mapping populations (Li et al. 2005; Prashar et al. 2014), and could be interpreted as the result of two closely linked genes or due to a pleiotropic effect of a single locus (van Eck 2007).

Haplotype structure; cryptic variation

Following on the association mapping we looked at the Linkage Disequilibrium (LD) patterns within this region. Using all markers from a 3.1 Mb region flanking the most significant QTL we observe a single haplotype block in which all significant markers trait associations were found, spanning a region of ~2.3 Mb. This large haploblock is not unexpected, as is postulated by Vos et al. (2016) that average LD is extensive in potato, where $D_{1/2}$ ranges from 600 Kb to 2.5 Mb in respectively old and recent varieties, suggesting that long haplotype blocks should be the norm in potato. In addition we observed that the significant marker-trait associations for eye depth are also present within this single large haploblock, suggesting the co-localization of both the *Eyd* (Li et al. 2005), and *Ro* locus within this region of 3.1 Mb.

To further understand patterns of variation among this haploblock we investigated the haplotype structure within the most significant markers, disentangling these 20 significant markers (Table 1) into 5 distinct haploblocks based on pairwise correlation between markers. Cofactor analyses suggest that within this haploblock we see cryptic variation in the form of multiple alleles at a locus. Nevertheless, a high correlation between SNPs could be due to other factors than presence of SNP alleles within the same haplotype. At this moment we cannot disentangle bi-allelic SNP information in their constituent alleles, limiting our understanding of the haplotype structure at this locus.

GWAS in potato

In this study we used Linear mixed models (LMM) with correction for population structure, as in previous papers, extensive population structure was acknowledged (D'hoop et al. 2010; Hirsch et al. 2013; Malosetti et al. 2007; Rosyara et al. 2016) (Vos et al. 2017). Using simple linear regression models resulted in strong *P*-value inflation, whereas using population-structure corrected using a kinship (K) corrected LMM effectively reduced confounding (see Figure 3) (Kang et al. 2008). Our analyses reaffirm the need for population structure correction to avoid an excess of false positive associations. This is also supported by Rosyara et al. (2016) where different Q (subpopulations) + K (kinship) model was used for structure-corrected GWAS. In their analysis the Q+K approach leads to the best structure correction, although the differences in performance between the Q+K model and the K -model are marginally. Strikingly the QTL on chromosome 2 is not observed using the 'general' model as used in GWASpoly (data not shown), but is observed using an additive model of trait variation. Previously Rosary et al. (2016) demonstrated severe loss of statistical power due to the loss of degrees of freedom, using the 'general' model, explaining the lack of power for QTL detection.

Missing Heritability – explained variation in phenotypes

Despite the occurrence of highly significant associations in the chromosome 10 region the highest explained variation for a single marker is 26%. Using a multi-locus stepwise regression with backwards selection we could explain up to 41% of the phenotypic variation including all significant markers within the chromosome 10 QTL region. Using the same approach with the top 20 significant markers we could only explain 36% of the variation in phenotype. In contrast, the broad-sense heritability (H_2) is estimated to be 0.8 (D'hoop et al. 2014), suggesting that a considerable part of the genetic variance is not explained in this study. A part of this so-called missing heritability might be explained by allelic heterogeneity (i.e. multiple alleles with similar or distinct phenotypic effects), which given the observation of different causal alleles for tuber shape is an likely explanation.

Indeed, the investigation of linkage disequilibrium (LD) patterns among between pairs of significant markers, identified five groups of highly correlated markers occur within

this region. At the moment we cannot conclude this, as phenotypic variation is explained not only by the alleles at this locus, but also similar alleles might have different effects depending on genetic background, potentially obscuring a marker-trait association.

Fine mapping of the major QTL on chromosome 10

In the potato germplasm the decay of linkage disequilibrium on average spans 5 cM (D'hoop et al. 2008, Vos et al. 2016), limiting the maximum resolution an association study can achieve. In our study we identified the most significant associations with tuber shape in a large haploblock of 2.3 Mb, where markers at the end of scaffold DMB385 showed association to tuber shape, but later fine mapping refined the locus to a region ~300 kb downstream (or south), but still within this haploblock. Within the 2.19 Mb region downstream of DMB385 only 13 additional markers were found, from which none in the 280 kb large candidate region on DMB546, suggesting an unfavourable marker density in this region.

Fine mapping allowed further mapping of the *Ro* locus to a region of 280 Kb, enriched for a cluster of cell wall III peroxidases. Initially within a region of 3.1 Mb a total number of 105 (female) and 27 (male) recombinants were identified, suggesting a unfavourable ratio of physical distance to genetic distance**.** In the second batch a total of six recombinants were found between marker Asp6678 and Per20801, although this 280 Kb region only spans a genetic distance of at most 0.62 cM in the female and 0.12 cM in the male parent. This limited genetic distance implies that a large population should be screened to find more informative recombinants to refine this interval further.

Candidate genes underlying the major QTL for tuber shape.

Within the candidate region of 280 Kb amongst 14 functional genes, a total of eight annotated peroxidase genes were located. Expression of other genes residing in this candidate region show that all other genes are not expressed, except one unknown protein (PGSCDMG20797). Repeat analysis identified three more peroxidase genes. Putative function of the peroxidase genes suggest that this peroxidase cluster is most likely to be involved in regulating tuber shape. Previously the involvement of peroxidases in organ size or shape was reported (Francoz et al. 2014, Passardi et al. 2005). In Arabidopsis peroxidases were found to influence root elongation (Passardi et al. 2006, Pedreira et al. 2011). Peroxidases are also found to be involved in shape variation within

the *Brassicaceae* genus, where enormous differences in organ morphology are explained by copy number variation in peroxidase genes in the pan-genome (Lin et al. 2014). The peroxidase cluster within this region is most likely to be candidate for modulating potato tuber shape, although none of the other genes can definitively be excluded. Subsequent studies using positional cloning and functional complementation are required to confirm the causative gene underlying this major effect QTL.

Multiple alleles

The association mapping suggests that continuous variation in potato tuber shape is almost exclusively regulated by the Ro locus, located in a region of 280 Kb on chromosome 10 for which multiple alleles were identified. Hence potato tuber shape is determined by multiple combinations of different causal alleles. Intuitively a continuous phenotype would imply a polygenic inheritance according to the infinitesimal model of Fisher (1918). However already in 1929 Sirks proposed that a single gene model with multiple alleles, each having a different effect, could explain quantitative trait variation. It might be argued that this strict Mendelian view of trait variation is an simplification, as trait variation cannot be seen independent from genetic background and environment, however the fact that tuber shape has a high heritability of 0.8-0.95 (D'hoop et al. 2008, 2011) suggests that the contribution of environmental factors is limited.

In our GWAS analysis we demonstrated the occurrence of cryptic variation at the *Ro* locus, suggesting the occurrence of multiple alleles for tuber shape, which was supported by Van Eck et al. (1994) and reaffirmed in this study. As potato varieties are characterized by a high heterozygosity, where many variants do have a low allele frequency (Uitdewilligen et al. 2013). Binary SNP markers will not capture all the allelic diversity occurring in the potato genepool (i.e. can only distinguish two alleles), and might be present on two or more alleles with similar or opposite effects (i.e. allelic heterogeneity) (Bergelson et al. 2009), leading to a reduced statistical power in association mapping (Schaid, 2004).

In addition to this, it is expected that when gene action is matched by the marker model the power of detection of QTLS is increased, which was demonstrated recently by Rosyara et al. (2016) using a simulation study. However identifying accurate gene action

models will be problematic for traits were multiple alleles contribute to trait variation, and when multiple QTLs with multiple alleles jointly explain phenotypic variation, limiting the use of these marker models in GWAS. For example, considering the alleles in the C × E experimental population we could postulate that tuber shape should have a dominant gene action (round over long). In contrast considering only round alleles, subtle additive effects can be found as two *Ro* alleles lead to compressed potato tubers.

Intriguingly, not only tuber shape displays the occurrence of multiple causal alleles contributing to trait variation. Previous research into other well-studied potato traits, like plant maturity and starch metabolism has identified multiple causative alleles at QTLs. Like tuber shape, differences in plant maturity are explained by allelic variation at a single locus, *StCDF1* (Kloosterman et al. 2013). Starch accumulation and metabolism has been found to be influenced by many genes, although only 15 QTLs were identified. Subsequent investigation of the identified QTLs revealed multiple causal alleles (Schreiber et al. 2014). Other examples are the GBSS I locus, determining the amount of amylose, where multiple causal alleles were identified (Van de Wal et al. 2001). Moreover starch phosphorylation in potato tubers is mediated by four genes (GWD, *SBEI SBEII, SSIII),* for which multiple alleles were identified (Uitdewilligen et al. 2013; Carpenter et al. 2015). Thus, trait variation might also be attributed to the occurrence of combinations of multiple causal alleles at a QTL position.

At the moment we do not know how many alleles for the *Ro* locus can be identified in the gene pool, nor their effects. Likewise we cannot estimate the effect of allelic heterogeneity, as we do not know the haplotype structure underlying these SNP markers. In the future the use of haplotype-specific markers (or combinations of these markers) will allow us to identify more adequately the underlying genetic architecture of tuber shape as well as improve the detection of functional alleles and their effect.

Acknowledgements

JHW is supported by a grant of the Dutch Science Organisation NWO (project 831.14.002). The potato breeding companies Averis Seeds B.V., HZPC Holland B.V., KWS POTATO B.V. and Meijer B.V. are acknowledged for contributing phenotypic data. The Dutch Technology Foundation (STW grant WPB-7926) financed the development of the SolSTW array and SNP-data production.

Additional files

Table S1: Phenotypic data of 537 varieties for potato tuber shape and eye depth

Table S2: PCR markers for recombinant analysis

Table S3: Tuber shape is significantly confounded with structure group

Table S4: Eye depth is significantly confounded with structure group

Table S5: backward elimination – multi-locus regression

Table S6: GAP-spanning BAC clones

Table S7: Genes (predominantly peroxidases) annotated in DMB546

Table S8: Results of fine mapping using recombinant screening.

Figure S1: High resolution melting analysis of marker Per20801 allows full genetic classification of C x E offspring in four Mendelian classes.

Figure S2: Distribution of tuber shape. Best Linear Unbiased Estimators (BLUEs) for tuber shape in a panel of 537 tetraploid potato varieties and advanced breeding clones, recorded on an ordinal scale (1=long, 3=long/oval, 5=oval, 7=round/oval, 9=round), display a normal distribution.

Figure S3: Distribution of eye depth. Best Linear Unbiased Estimators (BLUEs) for eye depth in a panel of 190 tetraploid potato varieties, recorded on an ordinal scale (4=very deep, 5= deep, 6=intermediate, 7=shallow, 8=very shallow), display a skewed distribution, as deep eyes are unmarketable.

Figure S4: Manhattan plot naive GWAS of tuber shape and eye depth

Figure S5: Co-factor corrected Manhattan plots of markers associated with tuber shape

Chapter 3

Exploiting short-read sequencing for the characterization of haplotype diversity in polyploid crops

Johan H. Willemsen, Jan A. M. L. Uitdewilligen, Richard G. Visser, Herman J. van Eck

Abstract

Background: In crops, haplotype identification is important for the identification of allelic diversity responsible for trait variation. For diploid plant species many approaches have been described that allow accurate reconstruction of haplotypes. However, in polyploids high-throughput haplotype detection is challenging due to computational complexity. There are several ways to obtain haplotype information in polyploids, e.g. read backed phasing, using next-generation sequencing (NGS) data. Sequenced fragments contain partial haplotype information. Exploiting this information allows the high-throughput identification of all haplotypes or alleles of a locus.

Method: We introduce an efficient algorithm to estimate haplotypes from sequencing data, without relying on previously obtained genotype calls. We divide the haplotype reconstruction into two steps: Short-range haplotype reconstruction, followed by haplotype extension.

Outcome: With simulated data we demonstrate that this approach leads to highly accurate haplotypes. We verify the approach with resequencing data of tetraploid potato varieties. As one of the first studies applying haplotype reconstruction in complex autopolyploid crops we highlight the challenges and limitations of using haplotype assembly for the reconstruction of haplotypes.

Key words:

Haplotype assembly, polyploid, sequencing, potato, haplotypes, allele diversity

Introduction

The identification of allelic diversity represents a daunting challenge for crop genomics, especially when polyploidy and heterozygosity interfere with genetic dissection of complex traits. So far, in polyploids such as potato (*Solanum tuberosum),* Chrysanthemum (*Chrysanthemum morifolium*), leek (*Allium porrum*) and alfalfa (*Medicago sativa*) genetic variation has been explored mainly as DNA variants in the form of single nucleotide polymorphisms (SNPs), multi-nucleotide polymorphisms (MNPs) and insertions and deletions (INDELs) (van der Geest et al. 2016; Yu et al. 2017; Uitdewilligen et al. 2013). These molecular markers address many questions related to marker assisted and genomic selection, association analysis and mapping of quantitative trait loci (QTL) in crops and model plants (Morrell et al. 2011).

Despite the success of single marker analyses, for dissecting complex trait genetics, the fundamental unit of inheritance remains the transmission of alleles or haplotypes. While bi-allelic SNPs will distinguish only two kinds of allelic variants out of many possible alleles, the joint observation of two phased bi-allelic SNPs will allow the discrimination between up to four different haplotypes. Therefore, an intuitive approach would be to use these haplotypes as replacement for single markers in genetic studies. It seems realistic to assume that the use of haplotypes will offer greater resolution in genetic analysis, for example by removing ambiguity from the association between SNPs and genetic effects, as is already observed in QTL analysis in a polyploid biparental population (Hackett et al. 2013). In genome-wide association studies (GWAS) this ambiguity is referred to as allelic heterogeneity (Bergelson and Roux, 2010). Allelic heterogeneity greatly affects the power to detect an association (Atwell et al. 2008; Schaid, 2004; Clark, 2004), but can be recovered when using multiple SNPs which are phased into haplotypes. The potential use of haplotypes is not limited to QTL discovery, because other studies suggest that substitution of markers by haplotypes allows a better estimation of identity-by-descent (IBD) at a particular locus (Meuwissen et al. 2000), which is important for marker-assisted selection.

While genome sequencing can rapidly identify allelic variation, it remains a challenge to identify haplotypes from sequencing data only. In diploids, considerable efforts were taken to develop approaches for reconstruction of haplotypes from sequencing data (Patterson et al. 2014; Aguiar and Istrail, 2012; Bansal and Bafna, 2008). Commonly

these approaches are referred to as haplotype assembly, whereas approaches that use statistical estimation of haplotypes, within large reference panels are referred to as haplotype inference (Browning and Browning, 2009). Despite the progress in haplotype assembly in diploids, haplotype assembly is lagging behind in polyploids. Up till now, haplotype information in polyploids such as potato, is obtained from laborious Sanger sequencing of (cloned) PCR amplicons, which is impractical for high-throughput haplotype reconstruction. Even, though many of these studies emphasize the need for haplotypes to get biological insight (Mosquera, et al. 2016; Schreiber et al. 2014; Uitdewilligen et al. 2012), no reliable and high-throughput haplotype reconstruction method has yet been developed.

Most autopolyploid crop species are highly heterozygous outbreeders, characterised by a high nucleotide diversity ($\pi = 0.02$ in potato; Uitdewilligen et al. 2013). Therefore short read sequencing technologies will produce reads that already comprise useful phasing information of sufficient variant positions to enable methods to obtain haplotypes. To assess the potential of haplotype assembly two questions need to be addressed. First, whether haplotype assembly is suitable to obtain high quality haplotypes in polyploids. Second, how useful this (partial) haplotype information is for application in routine genetic analysis.

Excellent methods for haplotype reconstruction have been published most methods have implemented the so-called MEC-formulation (Minimum Error Correction), which tries to minimize the number of conflicts between haplotypes and reads, i.e. trying to flip a minimum number of bases in the reads to obtain two haplotypes. In spite of the successful implementation of the MEC score (Schwartz, 2010), the minimization of the MEC score does not guarantee correct solutions in polyploids. Haplotype configurations could have equal MEC scores, notwithstanding that either one or more configurations are false (Aguiar et al. 2013; Berger et al. 2014). Hence, haplotype assembly software in polyploids need to implement different optimization criteria than only looking at the minimization of the MEC-score. The first approach proposed for polyploids, QualitySNP, was predominantly developed for EST data, and employed a heuristic assembly of sequences into haplotypes (Tang et al. 2006). The first approach suitable for NGS data, HapCompass (Aguiar et al. 2013), employed phasing of adjacent SNP pairs, and subsequent extending SNP phasings, while minimizing conflicts between the

read set and estimated haplotypes using an adaptation of the HapCompass framework (Aguiar and Istrail, 2012). Unfortunately, the authors of HapCompass also exposed that a single run will produce a haplotype assembly that is consistent with the data, but the assembly may not necessarily be the best haplotype assembly. HapTree (Berger et al. 2014) is one of the first haplotyping methods to use a rigorous statistical foundation, and framed the haplotype reconstruction in a maximum likelihood framework. To reduce computational complexity, HapTree uses a branching and pruning strategy to compute the most likely phasing through the first *m* SNPs. Subsequently it extends the haplotype by a single site and uses a relative likelihood function to determine which haplotype configuration is the most likely. When all SNPs are included, the most likely haplotype solution is reported.

Other approaches such as SDhap, use correlation clustering to obtain haplotype solutions (Das & Vidalo, 2015). Recently Xie et al. (2016) reported the development of H-Pop and H-Pop-G, which uses heuristic algorithms to obtain haplotype solutions.

A disadvantage of Haptree and Hapcompass is their sensitivity to genotyping errors. Both fully trust the SNP dosage, and therefore the haplotypes might be reconstructed incorrectly, as variant calling from NGS data will lead to a substantial number of erroneous genotype calls. H-PopG and SDhap do not depend on dosage information, but estimate dosages of SNPs during the assembly process by minimizing the MEC score using a novel heuristic algorithm. Based on simulations, both SDhap and HapCompass were shown to have inferior performance on short reads compared to H-Pop and H-PopG (Xie et al. 2016). Motezadi et al. (2017) compared Haptree with SDhap and HapCompass, and showed that Haptree performed best on their simulated data. It should be noted that the accuracy may very well depend on the type of (simulated) sequence data.

Here we set out to develop an alternative approach for haplotype assembly using NGS reads for the heterozygous tetraploid potato. Our main objective was to avoid the reconstruction of chimeric haplotypes, and to avoid reporting haplotype configurations that have equal support given the read assembly (i.e. reduce the number of switches, or equal MEC-score).

Our approach is based on a maximum likelihood approach using sequence alignments and variant positions as input, and based on early work in diploid haplotype assembly by Li et al. (1992) and Kim et al. (2007). Both these studies employed a divide-and-conquer technique for reconstructing longer blocks. In our method we start with reconstructing small haplotype blocks using a maximum-likelihood framework, over a sliding window, and extend haplotypes by joining them with a maximum-likelihood function constrained on the number of unique overlaps between blocks. Contrary to HapTree and HapCompass we do not require a priori genotype calls, but only need the positions of variants in the genome, for which during haplotyping the dosage is estimated.

Here we perform haplotype assembly using next-generation sequencing data of ~800 loci in 83 potato varieties (Uitdewilligen et al. 2013), which previously was used as SNP discovery panel for the development of a 20K SNP array in potato (Vos et al. 2015). So far haplotype structure in potato has only been investigated by phasing in structured populations (Bourke et al. 2015) or by reconstructing short range haplotypes based on Sanger sequencing data (Wolters et al. 2010; Schreiber et al. 2015). This study is among the first to use high-throughput sequencing data for the investigation of haplotype structure in polyploid species. We highlight the potential and challenges for haplotype assembly in polyploid crops. We demonstrate the performance of our approach for haplotyping with simulated data from Haplosim (Motezadi et al. 2016) and verify the accuracy of reconstructed haplotypes from real sequencing data of tetraploid potato varieties using pedigree relations between the varieties.

Methods

Data

Potato exome sequencing data

Data from targeted short read sequencing, captured with a SureSelect Bait library, representing 2445 target regions from 800 genes across the potato genome in 83 tetraploid potato varieties, were obtained from a previous project (Uitdewilligen et al. 2013). The total length of the adequately captured regions is 2.1 Mb (Table S1). The reduction of genome size facilitated the much needed read depth to achieve an average depth of 70× at reasonable costs (Table S2). Previously a total of 129K unique variants were detected across the complete panel, using FreeBayes (Garrison et al. 2009). For

further details about the bioinformatics analysis we refer to Uitdewilligen et al. (2013). Using these short reads we could test the performance of our method to reconstruct haplotype blocks within each separate contig.

Generation of simulated sequence data.
Arguably the real data described above, does not allow to explore all aspects influencing correct haplotype assembly. Therefore we simulated data with HaploSim (Motazedi et al. 2017) with variable

- insert size distribution (μ = 270, 470, 670 and 870 bp; $\sigma = 0.20 \times$ insert size),
- read length (100, 125 and 250 bp), and
- read depth (20×, 40×, 60× and 80×)

resulting in 4×3×4 = 48 parameter combinations. The error percentage was not simulated, but implicitly we used the known error profiles of HiSeq 2500 and MiSeq data (Huang et al. 2011). For each parameter combination, 100 regions of 5kb were randomly picked from the potato reference genome. Distances between variant positions and SNP allele dosages were randomly sampled from the frequency distributions as observed by Uitdewilligen et al. (2013), resulting in paired-end sequencing reads. Read depths above 60× could not be used to compare our method with earlier software, because HapTree could not handle this depth, as discovered before by Motazedi et al. (2017).

Performance assessment
To evaluate the simulated tetraploid haplotypes two criteria are used:

- the proportion of correctly reconstructed haplotypes within a haplotype block.
- the pairwise accuracy rate describes the proportion of correctly phased SNP pairs in a haplotype block.

In addition we evaluated the real data with two additional criterion:

- the proportion of isolated SNPs describes the proportion of SNPs that are not present in any haplotype block.
- the length of the haploblocks

Intuitively the first criterion, the proportion of correctly reconstructed haplotypes is the most appealing criterion, but this criterion is confounded with haplotype length. Longer

haplotypes are increasingly affected by haplotyping errors, leading to an overestimation of the error percentage, but large parts of long haplotypes might still be correct. In contrast, the pairwise accuracy rate will overestimate the correctness of haplotyping and underestimate the phasing errors when (1) not all alleles per locus are different and when (2) the different haplotypes are rather similar.

Validation with pedigree data

To benchmark the quality of haplotyping with real data, we could not compare our results to any reference, because the biological haplotypes are a priori unknown. However pedigree information informed us on the relatedness between potato varieties. Therefore the haplotype outcomes of related varieties can be mutually compared. In our dataset we identified the variety Markies and both parents Fianna and Agria. Each parent should contribute two alleles to variety Markies (ignoring double reduction), and the reconstructed haplotypes were compared with the pedigree. If we see haplotypes in Markies that differ from parental haplotypes, we know that at least one of the haplotypes is erroneous.

For the evaluation of concordance we compared haploblocks for those regions that are shared between the varieties, and calculated the number of absence/presence changes per haplotype, that is required to make the haplotype assembly concordant with the pedigree flow.

Haplotype assembly pipeline

An overview of the bioinformatics steps from data preparation to haplotype construction, hereafter referred to as the 'haplotype assembly pipeline' is illustrated in Figure 1. In short, after data preparation, for each region, alignments were extracted using SAMtools (Li et al. 2009) and per variant position the aligned sequencing fragments were parsed and only information regarding the variant position was retained (Step 1). Afterwards haplotype assembly was performed using a stepwise approach: First local haplotypes are reconstructed in a pre-set interval of n variants over a sliding window (Step 2). The local haplotypes are joined into longer haplotypes (Step 3).

Data preparation

For each region as specified in Table S1, read alignments were extracted using SAMtools and variant calling was performed with FreeBayes. These putative variants were further

filtered by using the vcflib library for parsing and manipulating VCF files. (https://github.com/vcflib/vcflib) As our haplotype assembly approach requires bi-allelic variants, multi-allelic SNPs were processed into multiple bi-allelic variants using the 'vcfbreakmulti' utility. Likewise complex variants, each composed out of multiple variant types, were parsed to their respective allelic primitives using 'vcftoallelicprimitives'. Due to ambivalence in read alignment the identification of INDELS from sequencing data is challenging. Although our haplotyping pipeline can process bi-allelic INDELS, we removed in this testing phase INDELS prior to haplotype assembly using the utility 'vcfsnps' from the vcflib suite.

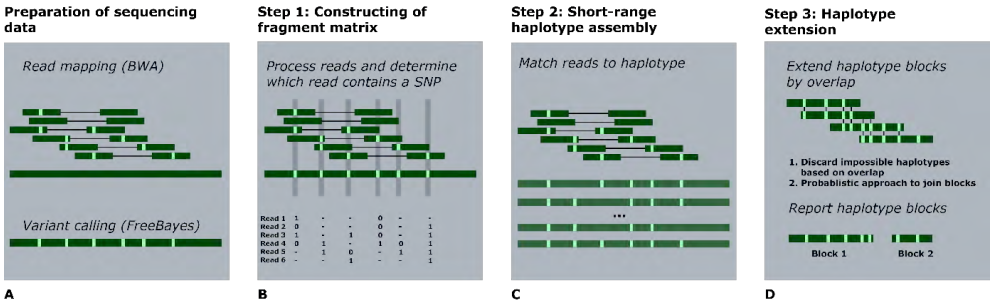


Figure 1. Overview of haplotype assembly pipeline: A) Bioinformatics steps involve read mapping using BWA, followed by variant calling with FreeBayes. B) Each read may contain multiple SNP(-alleles), which are scored. Only polymorphic positions are relevant for haplotyping. C) Reads are matched to local haplotypes, and the most likely short-range blocks are reported. D) Overlapping short-range haplotype blocks are used to reconstruct larger haplotypes. After haplotype reconstruction adjacent blocks are reported.

Step 1: Constructing of fragment matrix

To perform haplotype assembly within a variety, only information about polymorphic sites within this variety is needed. We extract sequenced fragments, aligned to the region under investigation, and determined for each fragment which SNP alleles are present and thus their phasing (Figure 1B). Each SNP allele is converted from molecular information into a binary system 0/1, where '0' always represents the reference allele, and '1' the alternative allele. Therefore, it is useful to represent the sequence alignment as a $m \times n$ matrix (Figure 1B), where m represents the number of polymorphic sites, and n represents the number of fragments present in the alignment.

Hence, at each region a $m \times n$ matrix is obtained where each fragment is described as read R_{ij} {0, 1, '-'}, where '-' denotes a gap before, after and in a paired-end fragment.

Furthermore, a '-' within a fragment also indicates missing information about any variant allele, and will only increase ambiguity in the haplotype assignment of fragments.

Step 2: Short-range haplotype assembly

Previous methods focused on finding haplotypes in such a way that the number of mismatches between fragments and predicted haplotypes is minimized (Anguir et al. 2013; Berger et al. 2014). In the same spirit we formulated an algorithm that reconstructs haplotypes and their dosages. This algorithm is inspired by early work in diploid haplotype assembly (Li et al. 1992; Kim et al. 2007) and uses the same basic assumptions (Churchill & Waterman, 1992), and is based on identifying the best match of sequencing fragments with a configuration of haplotypes.

Possible haplotypes

A k-ploid variety with k different alleles accommodates a set of haplotypes h_1, h_2 .. h_k. With n bi-allelic variant positions a total of 2^n haplotypes are possible. This results in a total set of phasings equal to the number of combinations of k out of 2^n with replacement. Constraining the set of potential phasings on the estimated SNP dosages does greatly diminish the number of possible haplotype phasings, but does increase the dependency on a priori computed dosage calls. We denote the haplotype k at variant position j by $h_{kj,}$ whereas the set of all possible haplotypes (2^n) is denoted by H. The set of possible phasing combinations are denoted by P, where phasing p is represented by P_p. Each phasing p contains a number of haplotypes that is equal to the ploidy level k.

Probabilistic model

We now describe a probabilistic model that incorporates PHRED-scaled sequencing error probabilities. For a single paired-end read the probability that a paired-end read (R_{ij}) originates from haplotype $h_{kj,}$ can be calculated from:

$$P(R_{ij} \mid h_{k,j}) = \prod_{j=1}^{n} P(R_{ij} \mid Y_{ij} = h_{kj})$$

Where j represents the j^{th} variant, and Y_{ij} represents the base quality at that position. Thus, the probability of observing fragment set R_{ij} given a set of haplotypes (phasing p_i) is:

$$P(R_{ij} \mid p_{\boldsymbol{p}}) = \left(\frac{1}{4}\right)^{m} \prod_{i=1}^{m} \sum_{k} P(R_{ij} | h_{k,j})$$

where ¼ represents the fraction of fragments emitting from each haplotype and k represents the number of haplotypes present in P_p. In words, we assume that ¼ of all sequencing fragments is produced by each haplotype, and we calculate the probability that a given fragment set originates from this particular combination of four haplotypes (Figure 1C). Subsequently we apply Bayes rule to obtain posterior probabilities of observing phasing P_p given the read data:

$$P\left(P_p \mid R_{ij}\right) \propto P(P_p)\, P(R_{ij} | P_p)$$

Here $P(P_p)$ is the prior probability of observing P_p, which is assumed to be equal and initialized as 1/ number of possible phasings). Consequently we can obtain the likelihood for the p^{th} phasing for the interval with length *j*. The likelihood of phasing P_p is:

$$P(P_p \mid R_{ij}) = P(P_p) \frac{P(R_{ij} \mid P_p)}{\sum P(R_{ij} \mid P_p)}$$

After calculating all likelihoods, the solution with maximum likelihood is reported. For each haplotype block, the confidence in each solution can be computed with the likelihood ratio between the two most likely solutions. If the two best solutions achieve similar likelihoods, so a likelihood ratio equal to around 1, and not exceeding the default odds-threshold of 50, we discard this solution from the output.

Step 3: Haplotype extension

The number of potential haplotypes increases exponentially 2^n with the number of variant sites considered. To reduce computational time, a stepwise approach (divide-and-conquer) is employed to assemble larger haplotypes. First, short haplotype segments of *n*-length are reconstructed over a sliding window. Each of these segments has a *n-1* overlap between adjacent haplotype blocks, which allows to use overlap between haplotypes to discard impossible solutions (Figure 1D). In practice, to reduce computation time, the default *n* is 4.

Frequently, situations will occur where overlapping haplotype blocks can be merged into an unique extension of the block length. If no unique haploblock extension is possible, then multiple haplotype configurations are tested with an equivalent likelihood function

as described above, to identify the most plausible extensions. Here we save computation time by constraining the possible set of phasings of overlapping haplotype blocks. In an iterative approach each haplotype block is merged with the adjacent block, until no merges can be made between overlapping blocks. The decision that no further merges can be allowed is based on the likelihood function or when all overlaps are processed. The final iteration of the haplotype block is reported (Figure 1D).

Implementation and availability

We implemented the haplotype assembly method in Python which can be downloaded from https://git.wageningenur.nl/wille094/Happy-haplotype-assembly/tags/v1.0 under the name Happy. Required formats for data input are for alignment data in BAM and variants specified by the VCF format version 4.1. The pipeline uses Numpy and Scipy packages as pre-requisites. Read information is extracted with SAMtools (Li et al. 2009), and processed reads are subsequently compressed with bgzip and indexed with tabix, whereas BEDtools (Quinlan and Hall, 2010) is used for intersection of variants in VCF files.

Haplotyping in 83 potato varieties

Genetic diversity

For each variety and each target region haplotypes were constructed. To describe the genetic diversity within potato varieties several statistics were calculated.

1. The number of unique alleles present in a reconstructed block per variety.
2. Nucleotide diversity, the average number of nucleotide differences per variant between two DNA sequences in all possible pairs of a single haplotype block of a single variety.
3. The distance of each allele to the reference genome of each homologous allele of a block.

For assessment of haplotype diversity among varieties we calculated the following statistics over a sliding window of *multiple* SNPs representing a length of 5, 10, 15, 20 and 25 SNPs. Only for segments in which more than 50% of the varieties provide haplotype data information (representing 41*4 = 160 allele observations), we report information about haplotype diversity/structure. Within a block of the frequency of each unique haplotype was determined. For each set of unique haplotypes, the nucleotide

diversity was calculated. The among-varieties nucleotide diversity was compared with the within-varieties nucleotide diversity.

Haplotype structure within the StGWD1 gene.
Previously haplotypes were obtained using sanger sequencing amplicons of one of the genes involved in phosphorylation of potato starch; the *StGWD1* gene (Uitdewilligen et al. 2012). Here we obtained the haplotype information by using sequencing-based haplotypes of two additional segments within the *StGWD1* gene (14 kb). For this we reconstructed the haplotypes across the 727 SNPs, and compared these across all varieties. For the purpose of this publication two segments were selected: 1) PotVar0091265 to PotVar0091285 comprising a region of 295 bp, and 2) PotVar0091399 to PotVar0091418 with length 248 bp. Within these segments the frequency of each unique haplotype was determined, and haplotype-defining SNPs were identified.

Results

Influence of sequencing parameters on haplotype assembly

To test our haplotype assembly method we used simulated 5kb regions with varying sequencing parameters, such as insert size, read depth and read length. This data was used to evaluate the variation in length of reconstructed haplotype blocks, as well as the accuracy of our haplotype assembly method. For all simulations we used the default settings of our pipeline.

Motazedi (2016) and Anguir et al. (2013) described the influence of insert size distribution on the length of haplotype blocks, where larger insert sizes will result in longer haplotype block length. For our method, we expect that local haplotyping (step 2) requires fragment spanning nearby SNPs, whereas haplotype extension (step 3) would require larger fragments to allow extension of haplotype blocks. Likewise read depth influences the ability to score dosages of individual haplotypes within each haplotype block. Therefore, we expect a relation between insert size, read length and also read depth, jointly influencing haplotype bock length.

First we explored the effect of increasing the average insert size distribution. Increasing the insert size in our simulations from 270 to 470bp did improve the length of a haplotype, in case of 100 bp reads, but a more substantial increase was observed for larger read lengths of 125 bp and 250bp. With increasing read depth we observed an increase

in haplotype block length, but only for low (20×) to moderate read depth of 60× (fig 2. ABC).

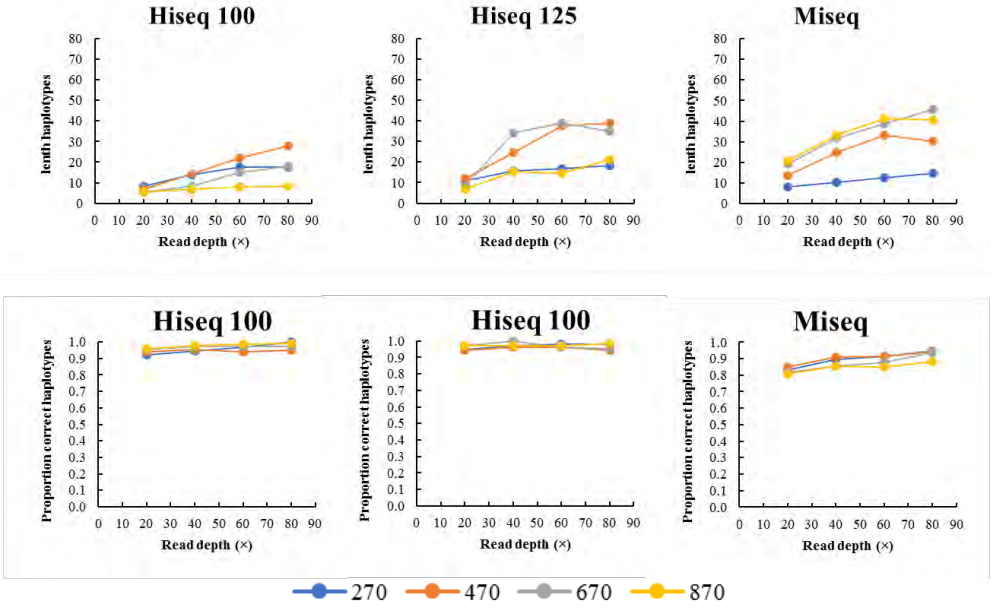


Figure 2. Evaluation of relation between read depth, fragment size and read length on haplotype block length and proportion of correct haplotypes. Horizontally total read depth is represented, and vertically the length of haplotypes in number of SNPs is visualized. The different colours refer to different insert sizes in bp.

Within each set of simulations we explored the accuracy of haplotype reconstruction by looking at the proportion of correct haplotypes. The accuracy over all simulations at low coverage of 20× was 90% ,and increasing read depth did increase the accuracy to 0.95 at read depth of 80× (Figure 2 DEF).

Overall the accuracy does not decrease substantial given an increased length of haplotypes, suggesting that long haplotype blocks are equally reliable as short haplotype blocks. In addition the trade-off between insert size, read depth and read length is an inevitable shortcoming of polyploid haplotype assembly. For the successful application of the divide-and-conquer approach, as used here, to reconstruct longer haplotypes, we would need reasonable read lengths (> 125 bp), but more importantly variation in insert size A large insert size would allow to 'bridge' over many SNPs, but lacks ability to phase nearby SNPs . On the contrary, a short fragment size can adequately reconstruct short-range haplotypes but fails to resolve long-distance phases.

Comparison with HapTree and Hapcompass

To benchmark the performance of our approach with HapTree and HapCompass, the two most commonly used software packages for polyploid haplotype assembly, we evaluated the reconstructed haplotypes using the same simulated data set as described above. Commonly haplotype reconstruction is evaluated by looking at the pairwise accuracy rate (Motazedi et al. 2017), although for short haplotypes the percentage of correct haplotypes might be a more practical metric. The objective of our method is to achieve high-quality haplotypes, and to avoid chimeric haplotypes, while putting less emphasis on haplotype length. Therefore we prefer the criterion of the proportion of correct haplotypes above the criterion of the pairwise-accuracy rate.

By using the proportion of correct haplotypes our method substantially outperforms other methods in terms of accuracy, irrespective of read length, read depth or insert size (Figure 3A). Previously it was shown that HapTree outcompetes HapCompass, and HapCompass was found to have equal performance over all read depths. Here we show similar findings, but using the proportion of haplotypes, HapCompass, only reconstructs 10% of the haplotypes without any error.

Strikingly, haplotype blocks reconstructed with HapCompass and HapTree are substantially longer than the haplotypes reconstructed with our approach, suggesting that the accuracy drops with increasing lengths (Figure 3B). Whereas our approach seems to counteract this by not allowing the extension of haplotypes, these other approaches do not adequately control this. Hence, although the complete haplotype that is reconstructed cannot be trusted (i.e. as judged by the proportion of correct haplotypes), locally these haplotypes might be correct. To illustrate this, we also determined the pairwise-accuracy rate to measure overall accurateness of the reconstructed haplotypes (Figure S1), which shows that HapCompass has a pairwise accuracy rate of around 0.67, irrespective of read depth and parameter setting, and HapTree at 0.85 at low read depths to 0.95 at a read depth of 80×. For our approach this ranges from 0.9 to 0.99 at 80×. We conclude that for both criteria our method outperforms HapTree and HapCompass, but notice that in the haplotype block we reconstruct are shorter than those obtained using HapTree and HapCompass.

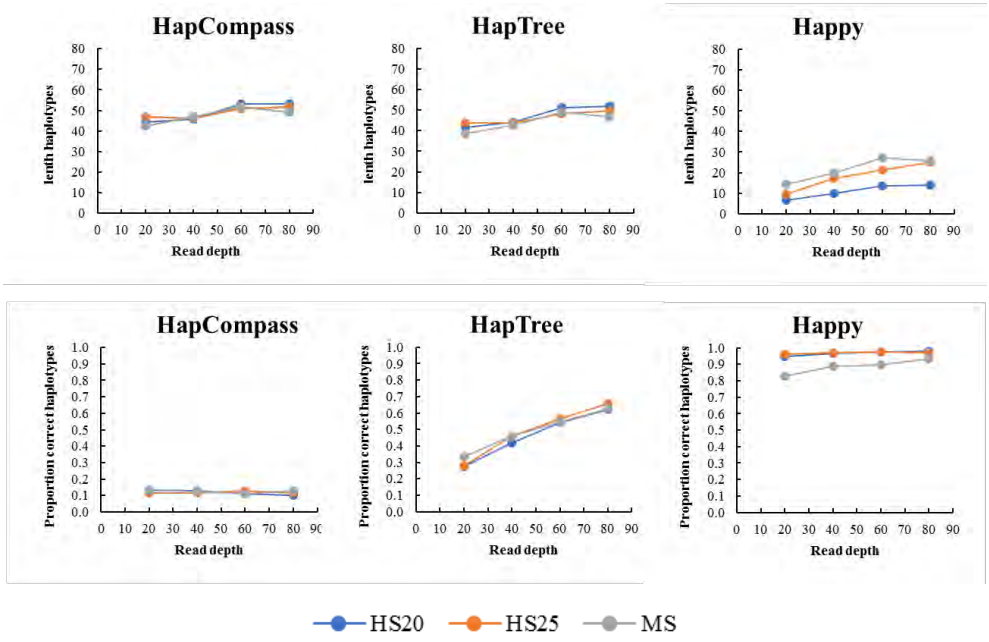


Figure 3. Results of simulation study with HiSeq (100 bp; 125bp) and MiSeq (250 bp). **A)** Our approach outcompetes both HapCompass and HapTree. **B)** Both HapTree and HapCompass show longer haplotype block lengths in numbers of SNPs. The colors refer to the technologies used (HS20: HiSeq 100 bp, HS25: HiSeq 125 bp, MS: MiSeq 250 bp).

Reconstructed haplotypes in potato

We used our approach for haplotype assembly on short read sequence data (100 bp PE; 270 average insert size) from Uitdewilligen et al. (2013) in 83 tetraploid potato varieties. Sequencing fragments were mapped to the reference genome, and after identification of target regions, a total of 2445 regions with sufficient coverage were detected. These regions have an average length of 874 bp, but also many regions with a size less than 200 bp were observed (Figure S1).

Within these targets, all variants were genotyped with FreeBayes, and initial filtering resulted in a total of 129,156 putative variants (Uitdewilligen et al. 2013). The total set of variants were filtered by read depth and genotype quality (GQ). To be selected, a variant required a minimum read depth of 15, and a genotype quality (GQ) higher than 26. As complex variants might have a detrimental effect on the haplotype reconstruction, only single nucleotide variants (SNPs) were retained. In the end we obtained 116.630 SNPs within the total of 2445 regions.

To reconstruct haplotypes, all 2445 target regions were individually processed and haplotypes were assembled. As most target regions are small, and considering our insert size distribution (average ~270 bp), we did not expect long haplotype blocks. Indeed, many of these regions were split into multiple haplotype blocks, and a difference is observed in the number of isolated SNPs, i.e. SNPs not present in any haplotype block. For all 83 varieties we observed an average haplotype block length of 413 bp (Figure 4), which was a 2.8× increase given the average insert size of 270bp, and considering the read length of 100 bp, a 4.1× increase.

In an ideal situation, all SNPs present within a sample are placed in a haplotype block. In reality this is often not observed, as a SNP can be adequately genotyped, but no haplotype-informative read pairs allow phasing of this SNP to adjacent SNPs. These SNPs are referred to as isolated SNPs. We expected that a higher read depth will decrease the number of isolated SNPs, as more haplotype-informative reads will be present at higher read depths. We determined which proportion of isolated SNPs are present within all 83 varieties (Fig 4A). Across all varieties, approximately 30% of the SNPs are not included in a haplotype block, and represent isolated SNPs. In addition, a strong correlation was observed with increase in read depth (Fig 4A). An increase in read depth translated to lower proportion of isolated SNPs, but with diminishing returns at read depth of 60×. Here within variety Pentland Dell, with mean read depth of 12.6×, almost 84% of the SNPs were not phased (=isolated). In contrast, in variety Kuras, with mean read depth of 152× only 12% of the SNPs were not placed in a haplotype block. Within the set of 42 samples sequenced with a read depth of 60× or higher, only 21% of the SNPs were not phased.

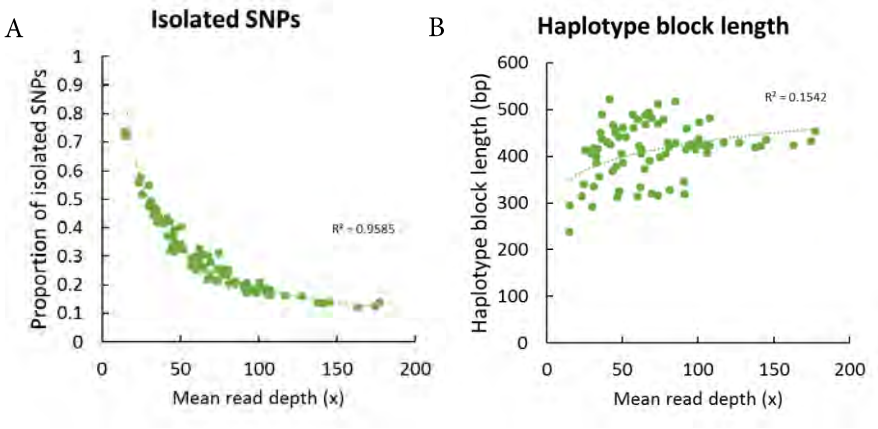


Figure 4. A) Correlation between read depth and the proportion of phased SNPs. B) Correlation between read depth and the haplotype block length.

To evaluate the relation between haplotype length and read depth we determined the average read depth at all SNPs within each haplotype block (Figure 4B). Here we observe that increasing read depth results in longer haplotype blocks, although read depth does not explain all variation in haplotype length. For example, in varieties like Nicola, haplotype blocks were on average 612 bp long, at read depth of 125-135, but the average haplotype block length was 512 bp (Figure 5).

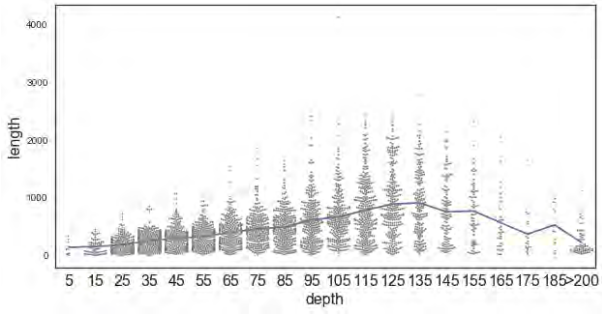


Figure 5. Example of read depth versus haplotype block length in variety Nicola. Each swarm represents a bin of 10×. The x-axis represent the average read depth haplotype block, related to the haplotype length in bp (y-axis). The blue line represents the average of each bin.

Validation with pedigree data

Although the above described simulation study suggests that our approach for haplotype assembly is reliable for a large variety of parameter settings, such as insert size and read depth, we also evaluated our experimentally obtained haplotypes by looking at concordance with pedigree data. For this we compared haplotype blocks across a

complete father-mother-child trio of parents Agria (P1) and Fianna (P2) and progeny Markies (Figure 6). Within the common blocks between these varieties, only of 1% SNP-alleles need to be flipped to make the haplotype configuration consistent with the allele flow through the pedigree structure. This suggests that the haplotypes are almost fully concordant with the pedigree structure.

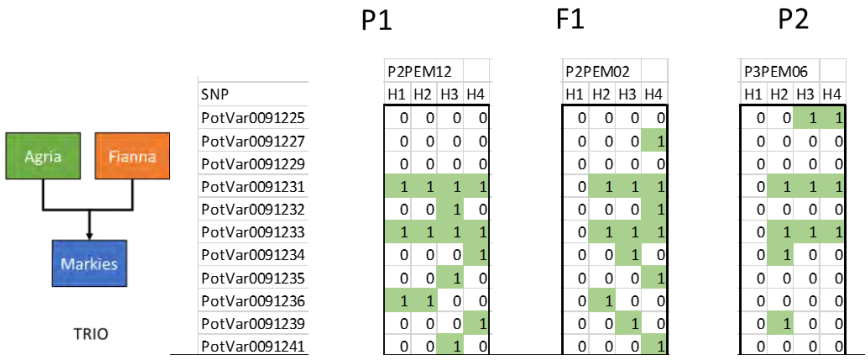


Figure 6. Example of a reconstructed block that is conflicting with the pedigree. A) pedigree B) Reconstructed block.in Agria, Fianna and Markies. Only one SNP-allele needs to be swapped to make this solution concordant with the pedigree, which translates in an edit distance of 1.

Consistency of haplotype assembly across varieties.

While our results show that it is possible to reconstruct accurate haplotypes within a single variety, we also want to compare haplotype blocks across varieties. A limitation of our data is that only exons are captured during sequencing, and genes might be only sequenced partly, resulting in multiple target regions that do not contain connecting sequencing fragments. In addition we also expect that haplotype blocks are not of equal size between varieties. For example a single haplotype block one variety might be split in two haplotype blocks in another variety. Such breaks between two adjacent haplotype blocks could occur because of several reasons: 1) Lack of read depth, 2) lack of overlapping reads between two adjacent blocks.

Here we explored if this comparison of haplotypes across varieties is possible using our reconstructed haplotype blocks. One notable effect of using sequencing data for haplotype assembly is the effect of read depth. We observed that haplotype blocks located in parts of targeted genes with systematic low read depth across samples are broken into

more separate haplotype blocks than regions that have a systematically high read depth, but also are more likely to show a break around that specific SNP with low read depth.

To investigate the consistency between all 83 samples, we used a sliding window approach we calculated the consistency of haplotype assembly by scoring how many varieties are fully haplotyped (i.e. all SNPs present in a single haplotype block). We expected that with an increase of this window in SNP numbers, the number of varieties that provide this information diminished. Indeed, while determining consistency over a sliding window of 5, 10, 15, and 25 SNPs, it was observed that the percentage of cultivars that provide full haplotype information diminishes from 84% to 61%, 50% and 40%.These segments represent an average length of 114, 225, 334, 435 and 531 bp.

An good example of lack of consistency across varieties is given in Figure 8, where haplotype blocks across the 83 varieties are visualized for the *StGWD1* gene. At low read depth (< 30×) comparison across genotypes is almost not possible, whereas high read depths across a set of consecutive variants allows a better comparison of haplotype composition across genotypes.

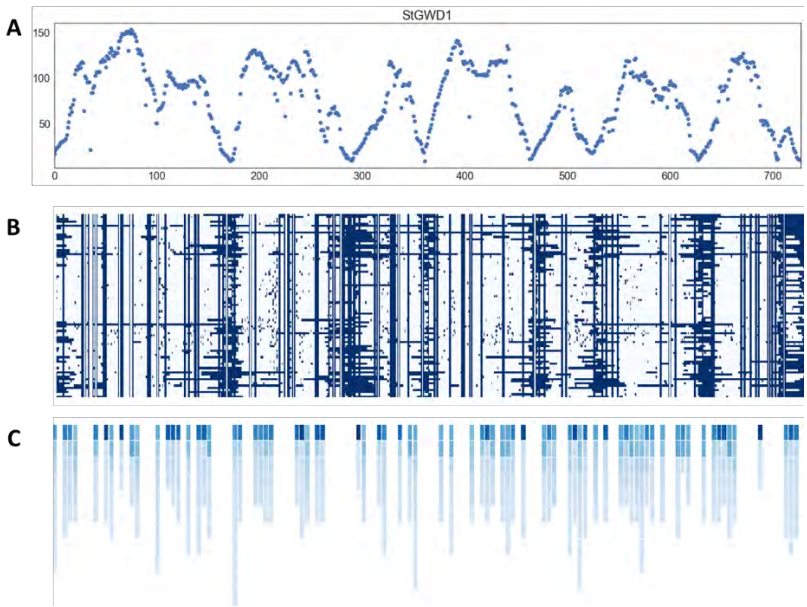


Figure 7. Consistency of reconstructed haplotype blocks across the *StGWD1* gene A) Read depth variation per SNP B) Fragmentation of blocks across varieties, where each row represents a genotype and the Blue colour represents locations at which two adjacent blocks are split, and therefore represent fragmented assemblies C) Frequencies of haplotypes discovered in blocks of 5 adjacent SNPs. The first row represents the most frequent haplotype, the second row the 2nd frequent haplotype, and so forth.

Haplotype diversity in potato

We next tried to estimate the haplotype diversity within the targeted regions. Previously it was determined that the variant density across varieties is around 1 variant on 42 bp (Uitdewilligen et al. 2013). To relate this to the reconstructed blocks we first calculated the average SNP density for each homologous allele of a haplotype block, to the reference genome. For all varieties, this resulted in 1 SNP every 94 bp of a homologous allele, ranging from 1 SNP every 84 bp in Pentland dell, to 1 SNP every 104 bp in Home Guard. In other words, if a single allele is selected, approximately every 92 bp a SNP is observed. This greatly exceeds the 1 out of 42 bp that is observed when looking over the four haplotypes.

As observed above the outcome of haplotype reconstruction is fragmented and cannot be easily compared among varieties, limiting our ability to explore the haplotype diversity across varieties. Therefore we first explored the haplotype diversity within a single variety. An average haplotype block in a variety contains 3.3 unique haplotypes, comprising 24 SNPs in a 413 bp region. Subsequently we performed an analysis over a sliding window of 5, 10, 15, 20 and 25 SNPs, recording the number of unique haplotypes within all cultivars, as well as the number of unique haplotypes across all cultivars. Strikingly the increase from 5 to 25 consecutive SNPs within a window, does increase the number of unique haplotypes from 7 to 15, but with diminishing returns (**Figure 8**). Both among-variety haplotype diversity as within-variety diversity exhibit the same diminishing returns.

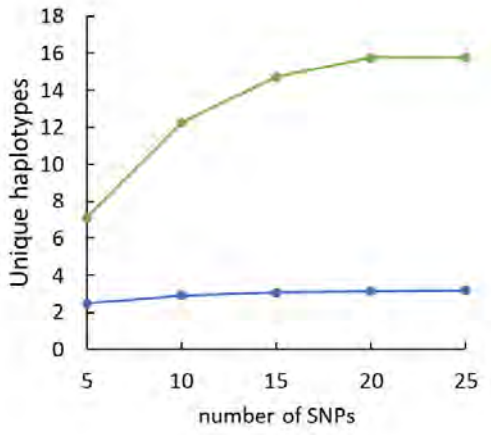


Figure 8. The average number of unique haplotypes in a single potato variety (blue) and the cumulative number of haplotypes among 83 varieties (green).

Case study: Haplotype analysis within the StGWD1 gene

To further explore the usages of the haplotype data presented in this study, we looked at two segments within the *StGWD1* gene, for which reasonable haplotype information is available. We tried to determine if our partial reconstructed haplotypes would allow to interpret the haplotype structure within the context of a candidate-gene approach. For the first segment (segment1) of 295 bp, we detected a total of 12 unique haplotypes, from which 8 are present more than once across all varieties (Figure 9). From the total of 18 SNPs only 14 are present among these 8 haplotypes, each of these haplotypes are named according to observed frequency. Among these 8 haplotypes, haplotypes F and G contain the most unique SNPs. Three haplotypes have a frequency higher than 0.2 (haplotypes A, B, C), and clearly represent common haplotypes.

We expected that these common haplotypes might represent old variation, whereas more recent introgressed alleles should be less frequent. Indeed these common haplotypes are already present in most old varieties, such as YAM (1787), VITELOTTE NOIR (1815) and BELLE DE FONTENAY (1885). When the reconstructed haplotypes are compared with the haplotypes from Uitdewilligen et al. (2012), haplotype G with, four haplotype specific SNPs is found to be co-segregating with haplotype H within amplicon Dex7 of Uitdewilligen et al. (2012), which previously was postulated to be originated from progeny of *S. demissum* introgression clone USDA 96-56. This clone was used as donor for R1 resistance against *Phytophthora infestans*. Likewise haplotype F co-segregates with haplotype E as discovered in Uitdewilligen et al. (2012).

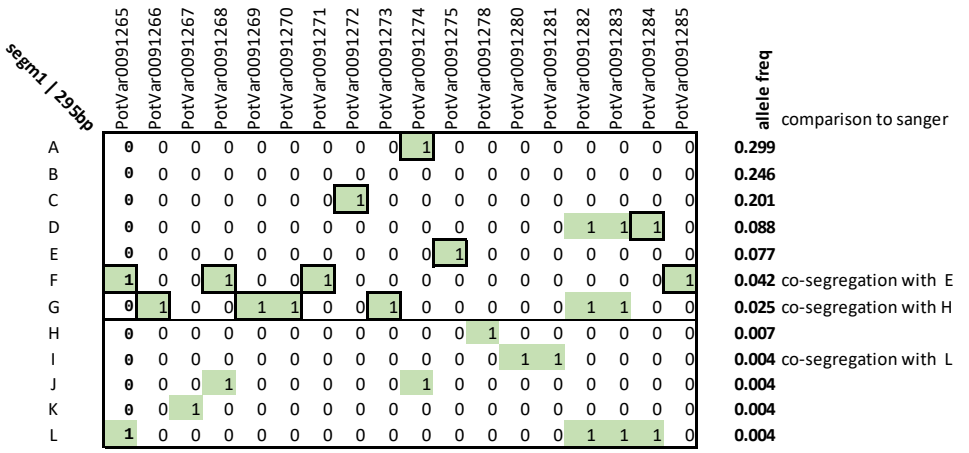


Figure 9. Phased GWD haplotypes of a segment of 295 bp. Haplotype-defining tag SNPs are bordered with black lines; Haplotype-defining SNPs are only defining in haplotypes that occur more than twice in the 83 varieties.

Haplotype frequency distribution over sliding window within StGWD1 gene

To investigate the haplotype structure at the complete gene we calculated for each adjacent block of 5 SNPs the number of unique haplotypes and their frequency (Figure 7C). Theoretically a block length of 5 SNPs would allow to distinguish between 32 haplotypes. Within the GWD gene we observe an average of 6.14 haplotypes for a segment of 5 SNPs. In general the haplotype frequency distribution follows an exponential distribution, which is expected when the population is a random subset of a random mating (data not shown).

Discussion

We developed an approach for haplotype assembly in polyploid genomes, which is able to reconstruct accurate haplotypes, even with limited sequencing data. So far no study has been published that makes use of haplotype assembly to reconstruct haplotypes in a of potato varieties, and only haplotype phasing in structured populations has been achieved using conventional linkage mapping (Bourke et al. 2016, Hackett et al. 2013) or by using a multi-point phasing approach, using a HMM-based approach in tetra-origin (Zheng et al. 2016).

In this study we present an efficient approach for reconstructing haplotypes from next-generation sequencing data. We validated our approach directly by using simulated data, but also indirectly by concordance to pedigree structure. Regardless of the simulated

dataset, our method outcompetes commonly used software HapTree and HapCompass. The strength of our approach is, that in contrast to previous methods, we use a divide-and-conquer approach to reconstruct haplotypes. The rationale behind our method is to focus on accurateness, and not on completeness of the haplotype assembly, as a single switch error among haplotype segments is not easily distinguishable from a novel recombination event. To achieve this objective, we first reconstruct accurate short-range haplotypes, and these (partial) haplotypes are used to reconstruct longer haplotypes.

Limits of polyploid haplotype assembly

To fully understand polyploid haplotype assembly we need to bear in mind the statements made by Aguiar et al. (2013), that given a number of SNPs, two or more completely valid phasings can be consistent with the sequencing data, but only sequencing fragments that connect the first and last SNP, may constrain the phasing solution to be unique (Figure 10). Thus, for any sequence of SNPs, a unique solution can be only guaranteed if paired-end reads span this whole segment. In a diploid organism, successful haplotype reconstruction only requires reads spanning adjacent SNPs, due to the uniqueness of phasings (i.e. observation of haplotype 00 means that 11 should be the alternative haplotype). As a consequence haplotype assembly in diploid organisms less challenging.

In polyploids we therefore need to consider, that if a part of a haplotype block contains a part that has less than four unique haplotypes, sequencing reads are needed that span across the full region (See Figure 10A**).** Indeed, in this study we show that the haplotype blocks reconstructed here contain an average of 3.3 unique alleles, hence regions of limited haplotype diversity are likely to occur frequently. As a consequence successful haplotype assembly in potato will require a long fragment size distribution to adequately bridge regions of limited haplotype diversity.

On the other hand, even when a region is entirely covered by sequencing reads, still multiple phasings might have similar read support, for example due to read errors, or allele imbalance during sequencing (Figure 10B). In those cases we cannot simply select one of the phasing that have equal read support, as it is just as likely that another haplotype solution is the true solution.

HapTree tries to reduce this problem by pruning haplotypes with lower likelihood, and reporting the haplotype that has highest likelihood, however if the resulting haplotype solutions have equal likelihood, a random solution is selected. In contrast, the HapCompass framework is based on reconstructing a number of paths through the so-called compass graph equal to the ploidy level (Anguir and Istrail, 2012), and report a valid phasing, that minimizes the MFER objective score. This valid phasing is not necessary the best phasing as noted by the authors (Aguiar et al. 2014).

We try to circumvent these problems by first constructing short-range haplotypes, followed by discarding (seed) haplotype blocks based on a likelihood ratio. In subsequent extension steps all reads connecting multiple short range haplotype blocks are used. This two-step strategy allows to 1) avoid regions of low-quality sequence alignment where haplotype assembly is likely to fail in every case, 2) avoid chimeric haplotypes by selecting only regions in which unambiguously haplotype dosage can be determined.

One of the advantages of this strategy is that efficiently short-range haplotypes can be computed as computational requirements are limited for short segments of only a few SNPs and subsequently during extension, in the worst case scenario only 35 possible phasing configurations have to be evaluated. Thus, a major advantage of this divide-and-conquer approach is that it allows to efficiently compute longer haplotypes.

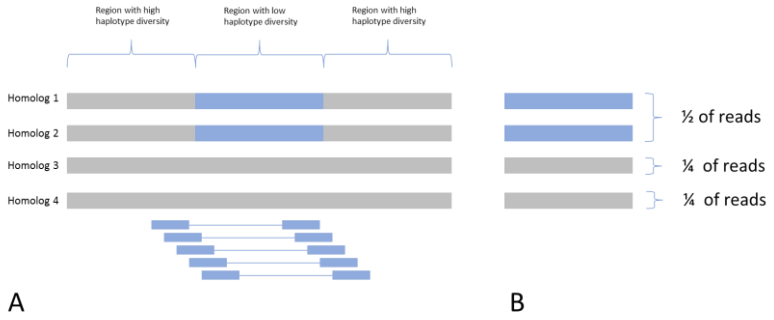


Figure 10. Limitations and challenges of haplotype assembly in polyploid organisms. A region of (partial) homozygosity is shown in blue. A) If a region of low haplotype diversity is encountered reads are needed that span that segment. B) If the haplotype diversity is less than 4, then additional freedom for dosage errors is there, and a high number of reads is needed to adequately estimate the dosage of the haplotypes.

Sequencing parameters influencing correct haplotype assembly

One of the goals for this study was to investigate which sequencing data is needed for adequate reconstruction of haplotypes. Using the experimental sequencing data we observed that that read depth is important for reducing the number of SNPs that are not phased (Figure 5A), and to a less extend influenced the haplotype length (Figure 5B). As no ground-truth haplotypes are available for this dataset, we supplemented these with simulation studies, evaluating a wider range of parameters such as insert size and read length; First a clear trend is observed that with increasing the fragment size, longer haplotype blocks are reconstructed (Figure 2ABC). Strikingly, reconstructing longer haplotype blocks not only required increasing the fragment size, but also requires an increased read depth, or length. This suggests that obtaining longer haplotypes is only possible if both short range connecting reads are present, in combination with longer fragments that span greater distances.

In practice a balance needs to be found between both read depth (> 40×), a broad insert size distribution, but also coupled with a reasonable read depth (> 125 bp). For example the short-read sequencing as used in this study, is sufficient to ensure short-range haplotype assembly because of high read depth (> 40×) and reasonable read length (100 bp), but the limited fragment size (270 bp average) clearly limits the length of the haplotype blocks.

Comparison to the HapTree approach

Our method bears most resemblance to the method of HapTree, which uses a relative likelihood function to estimate haplotypes (Berger et al. 2014). The HapTree approach constructs the most likely haplotype at two heterozygous SNPs, and subsequently extends in each iteration the haplotype by one SNP. As computing all phasings for an large number of SNPs is computationally extensive, a pruning strategy is used to remove phasings with low probability.

A disadvantage of this approach is that when the read depth or haplotype diversity limits the potential of haplotype assembly, easily switch errors can occur. The latter is also observed using the simulation data: While HapTree produces longer haplotype blocks than our method, the percentage of completely correct estimated haplotypes is low. In spite of this, the pairwise-accuracy ratio is substantially higher, suggesting that short-range haplotypes as obtained with HapTree are likely to be correct at higher read depths

(Figure S1). Previously Motezadi et al. (2017) showed that HapTree was most accurate in all scenarios, and that HapCompass has inferior performance compared to HapTree. Here we produced similar results, whereas HapCompass shows the most stable, but lowest accuracy of all three approaches.

Application of haplotype assembly for genetic analysis

One of the most promising applications for haplotype assembly in genetic studies, is the development of combinations of SNPs that are specific for any haplotype of interest. which can routinely be interrogated to trace haplotypes in a potato diversity panel. Here we tried to obtain haplotype-specific SNPs for the α-Glucan Water Dikinase (*StGWD)* gene. Previously it was found that two haplotypes had a positive effect on the amount of starch phosphorylation (haplotypes A+H) (Uitdewilligen, 2012) using conventional Sanger sequencing.

Using the partial haplotypes across varieties we detected a set of SNPs that would allow tracing these two haplotypes in a potato panel. We subsequently could extrapolate between the haplotypes as observed by Uitdewilligen (2011) and detect additional haplotype-specific SNPs suitable for use in marker-assisted selection, to improve starch phosphorylation content in potato breeding, which clearly shows the promise of using haplotype assembly for routine genetic studies.

From the total set of 12 haplotypes observed in a 295 bp window, a total of 3 common alleles are detected (> 5% MAF), cumulating to 75% of all observed haplotypes. The other 9 haplotypes are less frequent.

Haplotype diversity in the potato genepool

In addition to exploring haplotype structure across small segments of the StGWD1 gene, we also determined the number of unique haplotypes in these 800 genes, over a sliding window of multiple SNPs. Previously it was estimated based on patterns of linkage disequilibrium, that the potato genepool most likely contains around 6-10 commonly found haplotypes (Vos et al. 2015), which was also observed in the segments that were analysed within the StGWD1 gene (Figure 8C, Uitdewilligen et al. 2012). Other studies have shown that around 55 haplotypes are found within the *StCDF1* gene (N=58), whereas an investigation into the haplotype structure of the potato *ZEP* gene showed a total of 8 haplotypes. Here we extended this with an more comprehensive overview over

all 2445 target regions as used to reconstruct haplotypes. Within these regions we observe that a segment of 5 SNPs (~114 bp) generally contains 2.48 unique haplotypes, which shows a pattern of diminishing returns, and increases to a maximum of 3.19 unique haplotypes for a segment of 25 SNPs (~531 bp), which is within the bounds as observed by previous studies. Over all varieties approximately 7 haplotypes are found for a 5 SNP segment, which increases to around 15 haplotypes over 25 SNPs. This pattern of diminishing returns suggests that for adequate assessment of haplotype diversity of a locus, 25 SNPs should be sufficient.

Development in sequencing technologies

As suggested above, haplotype reconstruction in potato will require considerable amounts of sequencing data. In order to reconstruct longer haplotypes over further distances, our dataset could easily be complemented with a set of sequencing reads with large insert sizes, allowing to reliable reconstruct haplotypes across genes.

A viable alternative would be to use long-read technologies such as PacBio or Nanopore sequencing to reconstruct haplotypes. Although these 3rd generation sequencing technologies would certainly allow to increase the length of haplotype blocks, nonetheless, still considerable read depth is needed to either 1) determine the dosage of each haplotype, 2) accomplish enough overlap between reads, to allow haplotype reconstruction over larger distances than the average fragment length. At the moment these technologies have higher read error than conventional Illumina data, which needs to be compensated by increasing read depth as well. Nevertheless further development of these technologies might allow the reconstruction of gene-scale haplotypes with just a single read, making haplotype assembly less challenging. A promising new development is the chromium linked reads technology (10x), which would allow to have read pairs spanning 100 of kilobases. As the sequencing reads from this platform are conventional Illumina reads, these would potentially allow the reconstruction of complete haplotypes to a Mb scale, without much adjustment of conventional bioinformatics pipelines.

Prospect for haplotype assembly and challenges for further development

Although our method is suitable for reconstructing high quality haplotype blocks in a complex heterozygous genome of potato, we also noticed that comparison across varieties is challenging, as blocks are not uniformly present over all varieties. To improve the contiguity of our haplotype blocks we could potentially use imputation by imputing

partial haplotype blocks on the basis of complete haplotypes that are found in other varieties of the panel.

Recently in humans considerable efforts have been made to use haplotype assembly, but given the success of statistical phasing in diploids (Delaneau et al. 2014), one might argue research efforts of obtaining haplotypes in polyploids should shift to statistical estimation of haplotypes in large reference panels (Browning and Browning, 2008). However, so far only approaches have been published that allow phasing across limited sets of SNPs (Neigenfind, 2008; Su et al. 2008, Shen et al. 2016).

In general SNPs that are lowly frequent, may be phased reliable with using haplotype assembly, as rare SNPs provide difficultly for application haplotype inference (Snyder et al. 2015). Both approaches will provide highly accurate information at most SNPs, but the advantage of haplotype inference is that it allows phasing over long chromosomal regions, without the use of long fragments. Nevertheless, haplotype inference requires a reasonable panel size, whereas haplotype assembly can be performed in any single genotype for which reasonable sequencing information is available. In addition genetic relationships could be exploited to achieve pedigree-backed read phasing in pedigreed populations, such as bi-parental populations (Motezadi et al. 2018).

Conclusions

We have introduced a reliable method for haplotype reconstruction in potato and showed that this approach can successfully reconstruct haplotypes in tetraploid potato varieties. Here we aimed to answer two questions: First, whether single individual haplotype assembly is suitable to obtain high quality haplotypes in polyploids. The answer to this question is positive, but this depends on the length of haplotypes that are needed for downstream analysis, as current sequencing data has limitations for fragment size, read length and read depth, limit the reconstruction of haplotypes across larger regions. Noteworthy our approach has high performance, even when read depth and insert size is limited. The second question was how useful this (partial) haplotype information is for application in routine genetic analysis. We showed that we could reconstruct partial haplotypes for the *StGWD1* gene (Figure 10, 11), which directly can be used in marker-assisted selection, but also interrogated haplotype diversity across 83

potato varieties, showing that the results of haplotype assembly method such as the one presented here, might be fruitful for application in routine genetic studies.

Admittedly our method for haplotype reconstruction has still some limitations. In order to reliable reconstruct longer haplotypes a delicate balance needs to be found between read depth, fragment size and read length. This is not easily determined and might vary per crop species, or even between different genotypes of the same crop. This method can be improved by including population genetic approaches to allow multi-sample haplotype assembly. Our approach is mainly developed for the tetraploid potato, but suitable for all other ploidy levels, opening up avenues for other polyploid crops such as chrysanthemum, leek and alfalfa

Acknowledgements

JHW is supported by a grant of the Dutch Science Organisation NWO (project 831.14.002). The Dutch Technology Foundation (STW grant WPB-7926) financed the development of the SolSTW array and SNP-data production.

Additional files

File S1: Contigs used for haplotype assembly

File S2: Cultivar information and sequencing characteristics

Figure S1: Pairwise-accuracy rate of reconstructed haplotypes. The colors refer to the technologies used (HS20: HiSeq 100 bp, HS25: HiSeq 125 bp, MS: MiSeq 250 bp).

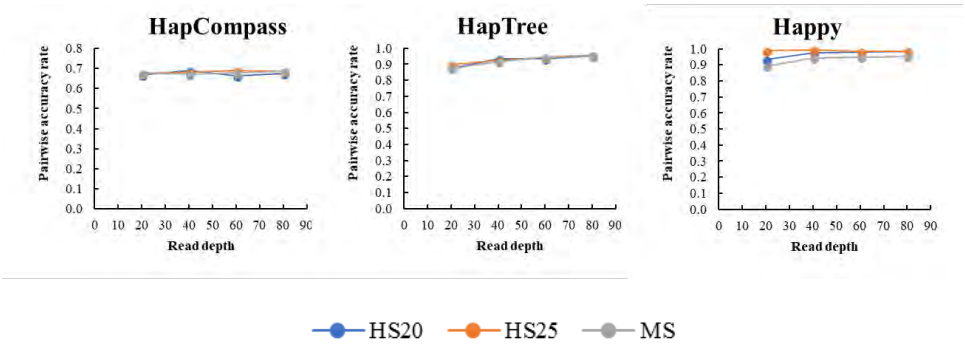
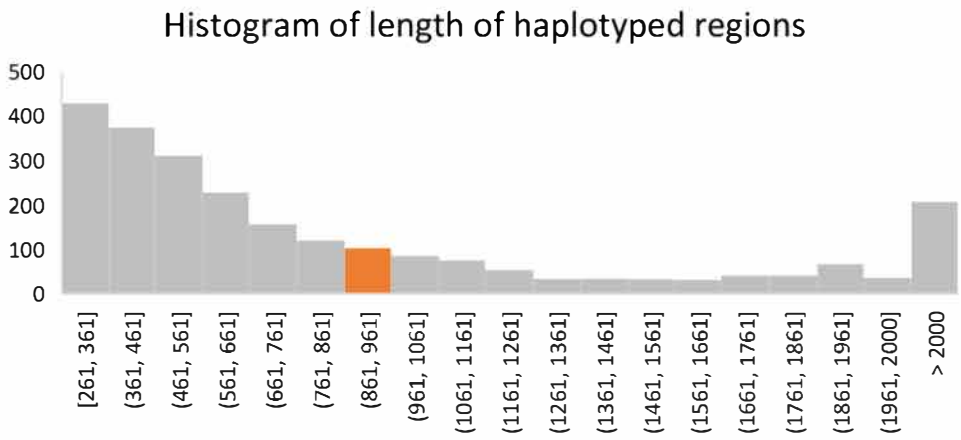


Figure S2: Histogram of interval lengths.


Histogram of length of haplotyped regions
500
400
300
200
100
0
[261, 361]
(361, 461]
(461, 561]
(561, 661]
(661, 761]
(761, 861]
(861, 961]
(961, 1061]
(1061, 1161]
(1161, 1261]
(1261, 1361]
(1361, 1461]
(1461, 1561]
(1561, 1661]
(1661, 1761]
(1761, 1861]
(1861, 1961]
(1961, 2000]
> 2000

Chapter 4

Haplotype inference in polyploid species and application to genetic analysis in potato

Johan H. Willemsen, Richard G.F. Visser & Herman J. van Eck

Abstract

Background: Identifying large numbers of bi-allelic SNPs can nowadays be achieved easily, but disentangling these into haplotypes is challenging as no linkage information between multiple SNPs is retained. To estimate these haplotypes, we developed a haplotype inference algorithm for polyploids. Our approach reconstructs haplotypes using un-phased genotype calls from GBS data or SNP arrays. Existing software for polyploid haplotyping, such as ShesisPlus, SATlotyper, and PolyHap, lack the ability to process the large amounts of SNPs present in highly heterozygous polyploid crops such as potato. The major improvement of our approach relies on a novel approach for joining short haplotype segments (estimated with the EM algorithm), allowing to scale haplotype inference to larger SNP numbers.

Results: Our results show that this approach is able to reconstruct high-quality haplotypes with a low number of phasing errors. Application of our approach to a dataset of un-phased SNPs derived from amplicon sequences demonstrated that most alleles could be reconstructed with high accuracy. In addition, we apply our algorithm on genotypic data from the potato 15K SNP array.

Conclusion: We present a scalable approach that accurately reconstructs haplotypes in polyploid crops. The resulting haplotypes are instrumental for analysing the haplotype composition of the potato gene pool, and haplotype based QTL discovery.

Key words:

Haplotype inference, polyploid, SNP, potato, haplotypes, allele diversity

Introduction

Genotyping platforms such as SNP arrays provide an efficient and accurate way to interrogate bi-allelic SNPs. In diploids it is sufficient to distinguish between homozygous or heterozygous genotype calls. In polyploids, SNP arrays should also assess allele dosage with high accuracy. Nowadays, high-density SNP arrays are available for genotyping of allele dosage in heterozygous polyploids such as potato, rose and chrysanthemum (Vos et al. 2015; Vukosavljev et al. 2016; Van Geest et al. 2017), and allows to characterize genetic variation in either bi-parental populations or association panels. These large datasets require the development of new tools able to handle the high marker numbers in polyploid genetic analysis such as linkage mapping (Hackett et al. 2008, Bourke et al. 2017), GWAS (Rosyara et al. 2016), and to facilitate the detection of marker-trait associations polyploid crops.

One of the limitations of genotyping data is the lack of information about linkage phase between SNP-alleles present across the many homologous chromosomes in polyploids, hereafter referred to as haplotypes or alleles. Whereas bi-allelic SNP arrays discriminate between two alleles at a single locus, haplotypes offer greater resolution to study the genetics in polyploids. For example, haplotypes instead of single SNP markers were used to detect marker-trait associations (Schaid et al. 2004, Buntjer et al. 2005). Another application would be the use of haplotypes in marker-assisted breeding in polyploid crops, which likely will improve the reliability of the marker-trait prediction.

Haplotype reconstruction in polyploids

So far a small number of successful studies have demonstrated the utility of haplotype information in polyploid crops, such as potato (Chapter 3, Bourke et al. 2017) and chrysanthemum (Geest et al. 2017), but also in diploid crops like maize (Huang et al. 2015). In most outbreeding polyploid crops, the determination of haplotypes is challenging, because haploids or homozygous inbred lines cannot be obtained easily. In general two approaches are used for haplotype reconstruction. Firstly, haplotype reconstruction can be performed by using physical linkage of variants jointly present in a single sequencing read (Aguiar et al. 2013, Berger et al. 2014, Motazedi et al. 2017), which will result in accurate haplotypes, but haplotype length is limited by sequencing characteristics such as insert size and read length (Chapter 3). Secondly, statistical methods can infer haplotype composition by exploiting SNP genotyping information

from a panel of unrelated varieties or from segregating bi-parental populations. These methods generally require moderate to large panels, but allow the phasing of SNPs into haplotypes over much longer distances.

From genotyping data to haplotype

The first challenge in polyploids, related to haplotype construction, is to achieve highly accurate SNP genotype calls, where fluorescent signal intensities from the SNP array are converted into an allele dosage. In diploids genotype calling is easy, because only three genotypic classes are expected (0 (AA), 1 (AB), 2 (BB), allowing to discriminate the heterozygous class from homozygous classes. In contrast, genotype calling in tetraploids needs to discriminate between nulliplex (AAAA), simplex (AAAB), duplex (AABB), triplex (ABBB) and quadruplex (BBBB). With higher ploidy levels, or with null-alleles even more genotype classes can be expected. Several R-packages have been developed for genotype calling using SNP array data in polyploid organisms, such as fitTetra and ClusterCall (Voorrips et al. 2011, Carley et al. 2017). When sequencing reads are used for genotype calling, other software such as Freebayes, QualitySNP or GATK can be used (Garrison & Marth 2012; Tang et al. 2006; McKenna et al. 2010).

Genotype calls obtained from sequencing data are error-prone, unless the read depth is adequately high. Previously it was determined that a read depth of 48× is necessary to obtain genotype calls with an accuracy of 95% (Uitdewilligen et al. 2013, Bastien et al. 2018), where at the moment most sequencing datasets fall short on this requirement. However in experimental data, read depth might vary between samples, resulting in varying error percentages for SNPs or between samples, suggesting that the accuracy of calls determined from sequencing data is substantially lower than the error rate of conventional SNP-array calls.

Progress in diploids

In diploids, many approaches have been published for haplotype inference using un-phased genotypes. One of the first approaches was introduced by Clark (1990), where the most parsimonious set of haplotypes consistent with the genotype data is estimated. Later, statistical approaches have been developed, assuming random mating, by using a probabilistic model to compute the likelihood of assignment of haplotypes to a genotype via the use of an Expectation Maximization algorithm (EM) or a Gibbs sampler algorithm (GS). Subsequent development of HMM-based models in Phase (Scheet and

Stephens 2001), FastPhase (Browning and Browning, 2007), Shapeit (Delaneau et al. (2008), and their widespread adaptation, suggest that HMM-based models are the most accurate to infer haplotypes. A likely reason why these HMM-based haplotype inference models are successful compared to regular haplotype inference models, is potentially due to better handling of erronous or the use of improved models for modelling haplotype diversity (Browning and Browning 2011). Nevertheless, others have questioned the added value of these methods due to similar performance of both likelihood based and HMM-based methods (Stephens and Donnely, 2003).

Polyploid haplotype inference

Haplotype phase inference in polyploids has been largely neglected by the research community. Zheng et al. (2016) describe a method suitable for tetraploid phase estimation in bi-parental populations, where an HMM-approach is used to first estimate parental homologs, and subsequently a probability is assigned whether this homolog is present in any given progeny (Zheng et al. 2016). In a F_1 population phasing is less complex, as only a maximum of eight unique homologs is expected, and hence only 36 genotypes (bivalent model) or 100 genotypes with the inclusion of double reduction.

In contrast, a panel with distantly related varieties might comprise a higher haplotype diversity, depending on the gene pool and composition of the panel. In the previous decade a few approaches for haplotype inference have been developed, but all lack the ability to phase larger numbers of markers into one haplotype. SATlotyper allows to reconstruct haplotypes by extending the Boolean satisfiability problem (SAT) approach to polyploids in order to solve the haplotype inference problem by minimizing the number of haplotypes explaining all SNPs (Neigenfind et al. 2008). Another package 'PolyHap' applies an HMM-approach to infer ancestral clusters for each haplotype, and subsequently assign an ancestral allele to each individual (Su et al. 2008). A more recent contribution is ShesisPlus, which uses a partition-ligation EM- algorithm to compute haplotypes in polyploids (Shen et al. 2015).

Aim of this study

We aim to improve the methodology for haplotype estimation, and use much larger numbers of SNPs to reconstruct haplotypes. Such haplotypes may span a large (e.g. gene sized) interval in species with a high nucleotide diversity. To achieve this, we developed

a heuristic algorithm that constructs longer haplotypes using a divide-and-conquer strategy. This strategy joins short haplotype segments that are estimated using a naïve EM-based haplotype inference. We performed a validation study using a set of haplotypes for the *StGWD1* gene as obtained with Sanger sequencing. In addition, we performed haplotype inference in a panel of 537 potato genotypes to explore the haplotype composition in intervals with high marker density. Our results show that the potato genome is characterized by few common and many rare alleles.

Material and methods

Genotype datasets

Data grouped into SNP dense intervals: A set of 537 tetraploid varieties was genotyped using the 15K SNP array (Vos et al. 2015). Details on the composition of the variety panel can be found in D'hoop et al. (2008) and details on the SNPs in Vos et al. (2015). The array includes dispersedly located SNPs from the SolCAP array (Hamilton et al. 2011), along with densely clustered 'PotVar' SNPs, which were obtained after a targeted resequencing of 807 genes (Uitdewilligen et al. 2013). Due to the physically uneven distribution of SNPs we could define SNP dense intervals. To delineate these intervals a distance cut-off between adjacent SNPs was used. A new interval is defined when the next SNP is at >10 kb distant from the previous SNP. For all intervals the SNP calling data was used as input to reconstruct haplotypes.

Validation data: Sanger sequences of two amplicons of approximately 0.6 kb were available from the *StGWD1* gene across a variety panel of 430 clones (Uitdewilligen et al. 2013). The resulting dosage calls (file S1) of 78 sequence polymorphisms were used as input data for haplotype reconstruction with our algorithm. For multi-allelic SNPs all alternative alleles were grouped. From these 78 SNPs three show complete linkage disequilibrium. Manual reconstructions (Uitdewilligen et al, 2013) resulted in 16 haplotypes with allele frequencies ranging from 31.8 to 0.1%, whereas here haplotypes A_1, A_2, A_3, A_4 are grouped into one haplotype (A), resulting in 12 haplotypes. Across the 430 potato varieties, four haplotypes per variety were assigned (file S2). The haplotypes shown in file S2 can be seen as ground-truth haplotypes and were used to benchmark the performance of our algorithm.

Problem formulation & implementation

Description of the algorithm

Our two step approach (Figure 1) consists of (1) linkage phase estimation between all SNP pairs using the EM algorithm described by Excoffier and Slatkin (1994), and (2) stitching these pairwise phasings into full-length haplotypes with a new iterative algorithm. The method can phase genotype calls either from SNP arrays as well as from sequencing. Genotype calls from sequencing will only reach the required data quality when polyploids are sequenced at great read depths.

Problem formulation

The genotype of a bi-allelic SNP, with alleles 0 and 1, indicating the reference and alternative allele, is described by the dosage of the alternative allele. In tetraploids the SNP dosage is defined as the sum of the alternative alleles (0, 1, 2, 3 and 4), and is referred to as nulliplex, simplex, duplex, triplex and quadruplex, respectively. With a set of n bi-allelic SNPs the total possible haplotypes is 2^n haplotypes. At ploidy level k every individual may carry between 1 up to k possible haplotypes out of 2^n, ignoring null-alleles.

The aim of haplotype inference is to find the set of 1 up to k haplotypes that best explains the individual SNP dosages within each individual. Phasing of a SNP in a haplotype is unambiguous if the SNP is homozygous (dosage = 0 or k), or when this SNP is the only heterozygous locus. Between two or more heterozygous SNPs, the inference of the linkage phase requires estimation. In polyploids 'linkage phase' needs to be specified across multiple haplotypes. Coupling phase refers to linkage between SNP alleles belonging to one haplotype, and implies multiple repulsion phase linkages with the SNP alleles at other haplotypes. In this paper we avoid coupling and repulsion, but use the term 'linkage phase' to indicate one connection between two SNP alleles in one haplotype, being a 00, 01, 10 or 11 connection.

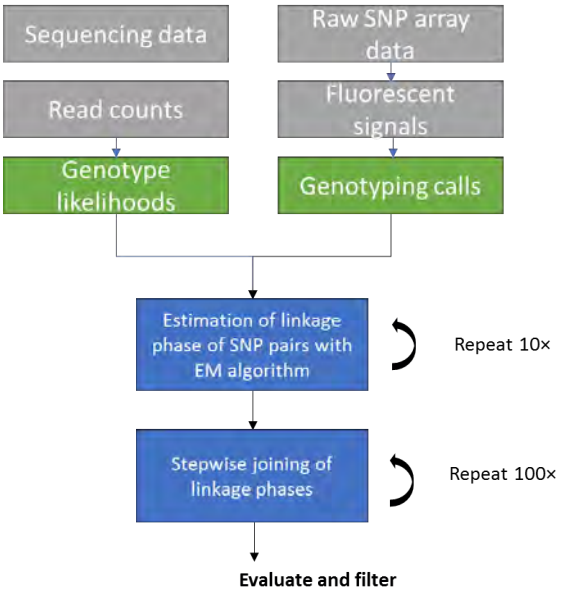


Figure 1. Schematic overview of the haplotype inference method, which uses as input for phasing genotype calls from either re-sequencing data or genotype calls from SNP arrays.

Method description

Pairwise phasing

When assuming random mating, the Hardy-Weinberg equilibrium (HWE) can be expected. At HWE the probability of observing a genotype is equivalent to the product of underlying haplotype frequencies (Excoffier and Slatkin, 1995). Hence, an estimation of the haplotype frequencies would allow computing the probability of observing a particular haplotype combination in a certain individual. The problem of estimating haplotype frequencies, and determining the linkage phase of the SNPs in each individual can be solved by applying the EM-algorithm, as was shown in diploids (Excoffier and Slatkin, 1995), and generalized to polyploids by Shen et et al. (2016). Here, the EM algorithm as proposed by Shen et al. (2016) was used to compute the pairwise linkage phases between all SNP pairs. In our dataset missing data were imputed by assigning an equal probability to any marker dosage. When sequencing data are used, our implementation of the EM-algorithm does allow to use not only discrete genotype calls, but also genotype likelihoods.

Each run of the EM algorithm on a randomly drawn subset (e.g. 50%) of the individuals returns different haplotype estimates, because of different haplotype frequency estimates. Usually ten runs of the EM algorithm on subsets of individuals is sufficient to distinguish between pairwise linkage phases that are robust to subsampling. Only those linkage phases that were observed in the majority of sampling runs are used as input for the joining step.

Joining step

Although, the EM-algorithm allows to compute linkage phases for more than two SNPs, the downside of this algorithm is, that computation time scales exponentially with the number of SNPs included in a haplotype. Although this can be sped up with using approximations of the EM-algorithm (Shen et al. 2016), in practice haplotype reconstruction with a moderate number of SNPs (e.g. 10 - 20 SNPs) requires enormous computational resources. To circumvent this, we employ a stepwise strategy, which uses the robust pairwise linkage phases to determine the haplotype composition in one individual.

Consider a set of three SNPs in a tetraploid individual, then $2^3 = 8$ putative haplotypes are possible, resulting in a total of 330 possible combinations of haplotypes, whereas only four haplotypes can be expected to be true. These four haplotypes are inferred by joining the EM derived linkage phase information of three pairs of SNPs as shown in Figure 2. Often there is not one solution when joining two SNP pairs, but the triangle of three SNP pairs will always deliver a unique solution. Erroneous haplotype solutions do not result from this joining step, but from an erroneous linkage phase estimate in one of the underlying SNP pairs.

Figure 2 assumes three linkage phase estimates (00, 01 10, 11), (00, 10 01, 11) and (00, 00, 11, 11) between SNP pairs 1-2, 2-3 and 1-3. The unique solution (000, 010, 101, 111) is not yet obtained by extending linkage phase information of SNP pair 1-2 with 2-3, but only after the information of pair 1-3 is integrated (Figure 2C). Also, the linkage phases between SNP2 and SNP3 are congruent with the solution. In principle one unique solution is expected (e.g. all linkage phases are in agreement with the set of four haplotypes), but when linkage phases result in equal support for multiple solutions, then one solution is reported that has minimum mismatches.

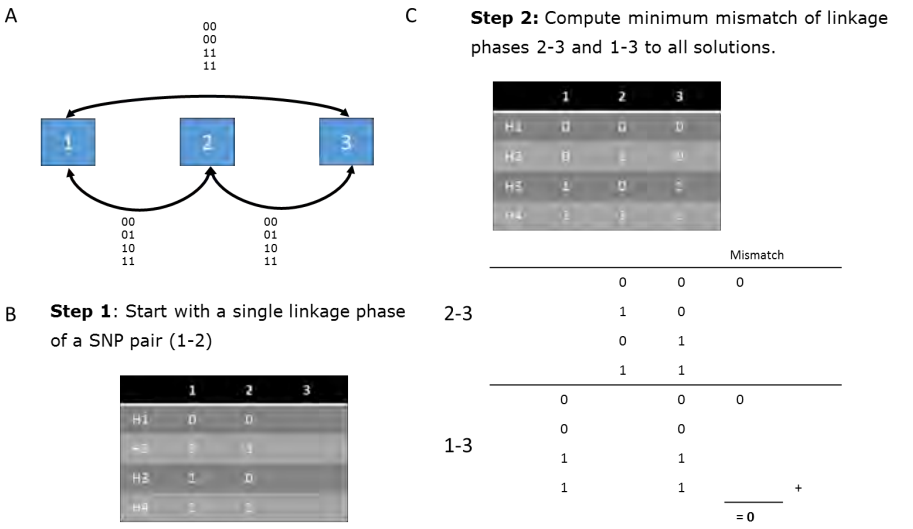


Figure 2. Joining of linkage phases into full-length haplotypes A) Three duplex SNPs are visualized as nodes, whereas edges represent the EM estimated linkage phases between these SNPs. B) In the first step a single arbitrary linkage phase is selected (in this case SNP pair 1-2). C) This haplotype of 2 SNPs is extended with a single SNP. To select the best solution the number of mismatches is calculated of all linkage phases of this SNP (2-3, 1-3). The solution with minimum mismatch is reported.

Implementation

To reconstruct full-length haplotypes we first select a random seed-SNP from an interval. In each subsequent step the haplotype solution is extended with one randomly chosen SNP. Each addition results in up to 4! = 24 possible linkage phases between alleles (or haplotypes) between the previous SNP (or haplotype) and the newly added SNP. A mismatch is a contradiction between a linkage phase estimate and a reconstructed haplotype. Mismatches may arise due to errors in genotype calling and/or EM estimation of linkage phases and result in building of incorrect haplotypes. Minimisation of mismatches is a step allowing the selection of the best solution out of the possible linkage phases.

By adding a single SNP in each extension step, we minimize the computational complexity caused by the large number of solutions in long haplotypes. More precisely, this stepwise method evaluates $24n$ solutions for haplotypes of n SNPs, where otherwise $(2^n+3)!/4!(2^n-1)!$ solutions are to be considered, assuming tetraploidy.

Because the joining algorithm is input-order dependent, we might obtain slightly different haplotypes after each run. When the algorithm is run multiple times (e.g. 100

times, but proportional to the number of SNPs), we assume that the most frequently observed haplotyping solution is the most likely haplotype configuration given all pairwise EM phase-estimates. To quantify the likelihood of this best solution, we calculate the ratio between the frequency of the best and the 2nd best solution. This 'Phase-ratio' allows to discard potentially unreliable haplotype before downstream analyses. We recommend to discard solutions with Phase-ratio less than 5.

Evaluation of haplotyping solutions

To compare inferred haplotypes with ground-truth haplotypes several measures are available to assess the performance of our software. These measures assess different aspects of the quality of the haplotype construction and cannot be directly compared.

- **Reconstruction Rate:** Measures the proportion of haplotypes that are correctly estimated (Motazedi et al. 2017).
- **Switch Error:** Measures the proportion of heterozygous SNPs whose phase is wrongly inferred relative to the previous heterozygous site (Neigenfind et al, 2008). In contrast to Reconstruction Rate the Switch Error is not inflated, but scales proportional with increasing haplotype length.

Availability of software

The procedure is available, implemented in a set of Python scripts, whereas both pairwise phasing (EM) as subsequent joining procedure are used to reconstruct haplotypes. As input a matrix with marker scores on rows (0-*k* ploidy), and columns corresponding to varieties. The scripts can be found on the Gitlab repository located at: https://git.wageningenur.nl/wille094/Happy-haplotype-inference/tags/0.8.2.

Results

Haplotype reconstruction in StGWD1 and algorithm validation

We developed an approach for haplotype inference which first estimates two-locus SNP linkage phases using population data. Subsequently, these linkage phase estimates are used to join SNP alleles into full-length haplotypes within an individual. For validation, dosage calls of 78 sequence polymorphisms (file S1) from two PCR amplicons of approximately 600 bp length were used. Our software detected a total of 24 haplotypes

across 380 varieties, of which eight were found more than five times in the whole dataset. The complete results are shown in file S3. When comparing our inferred haplotypes with the 12 ground-truth haplotypes (Uitdewilligen et al. 2013) ten haplotypes were correctly reconstructed, and correctly assigned to the varieties, and with the correct allele dosage. These included six haplotypes (A, B, C, D, E, F) with an allele frequency above 5%, as well as four out of six haplotypes with an allele frequency below 5% (G, H, J, K). Only haplotypes I (0.2%) and L (0.1%) were not reconstructed. Our software reconstructed eight erroneous haplotypes with allele counts of 1-4, (e.g. MAF of 0.1% to 0.25%) often chimeric due to incorrect linkage phase estimates.

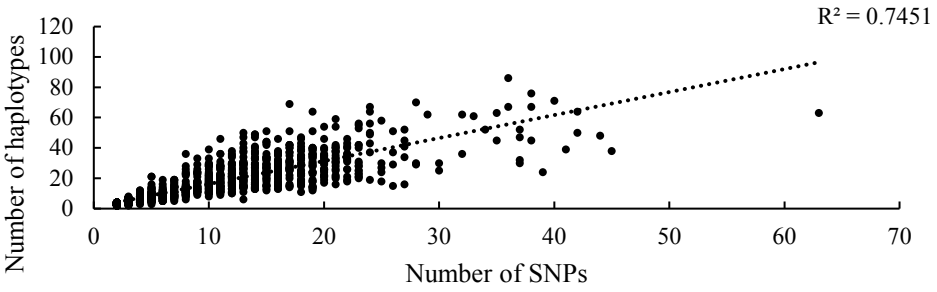
The accuracy of the haplotypes as judged by the reconstruction rate was 98%, resulting in only 56 cases with haplotypes that were erroneous, but these errors were distributed among 14 varieties. A more stringent setting of the threshold of the Phase-ratio parameter, from 2-5, resulted in an increase of accuracy, and at Phase-ratio of 5 resulted in a near perfect reconstruction rate (0.998). The disadvantage of a more stringent Phase-ratio of 5 is that approximately 10% of varieties are not inferred. The calculated switch error, indicative of chimeric haplotypes, was 0.45%, and is predominantly contributed due to chimaeras and not due to dosage errors. The reconstruction of haplotypes from these 78 SNPs in 438 varieties took only 5 minutes on a conventional desktop computer, suggesting that we can proceed with our genome wide high-density SNP datasets.

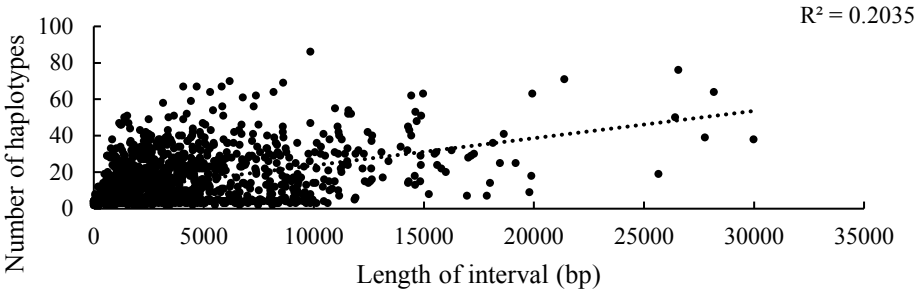
Haplotype reconstruction in tetraploid potato

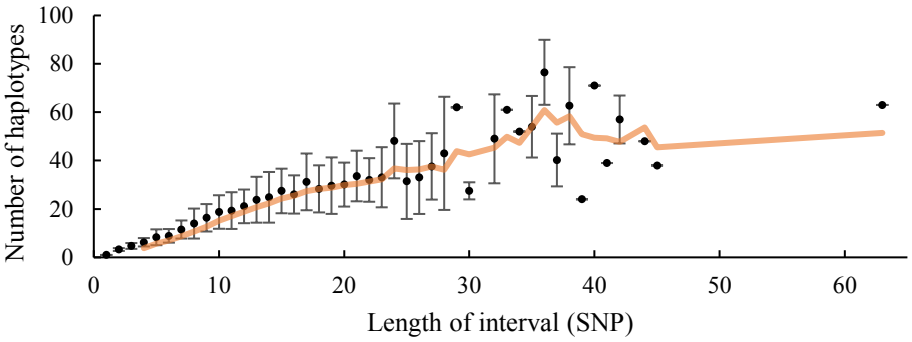
From the above-presented validation of our haplotype inference approach, we conclude that the inference approach is sufficiently reliable and can be used to detect haplotypes. Subsequently we applied this method on SNP data originating from a potato variety panel. Before phasing, the 14389 SNP markers were divided into 3217 intervals with high marker density and spaced <10kb. From the total of 3217 intervals, 1738 containing a single isolated SNP, and did not require haplotype reconstruction (File S4). The average number of SNPs within the remaining 1479 intervals is 8.4, ranging from 2 to 63 SNPs. The average length of an interval is 3217 bp, ranging from 1 to 29978 bp. Within the 1479 intervals, we estimated the phasings of SNP pairs and performed haplotype inference by using these pairwise phasings as input.

Subsequently, for each interval we calculated the haplotype diversity in number of alleles and allele frequency of each allele. As expected from the exponential frequency spectrum

of the SNPs, our haplotype reconstructions resulted in a similar distribution of haplotype frequencies. Most loci have a moderate number of common haplotypes (MAF ≥ 0.05) and larger number of rare haplotypes. The average number of common haplotypes is 3.6 per locus, ranging from 1 to 8 (Figure 3C). The cumulative allele frequency of common haplotypes is usually high, and only starts to drop for intervals with large numbers of SNPs (File S4). On average across all loci the common haplotypes cumulatively explain 92% of the allele diversity.


R² = 0.7451
Number of haplotypes
Number of SNPs


R² = 0.2035
Number of haplotypes
Length of interval (bp)


Number of haplotypes
Length of interval (SNP)

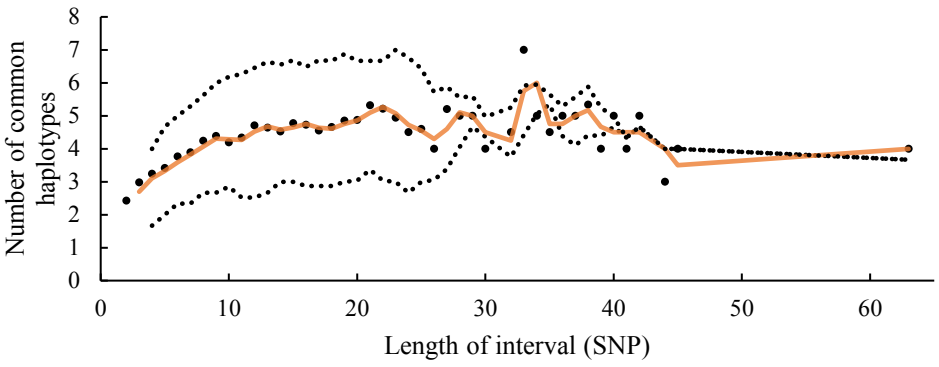

Number of common haplotypes
Length of interval (SNP)

Figure 3. Correlation analysis between haplotype length and haplotype diversity. A) Correlation between number of SNPs and number of haplotypes. B) Correlation between physical distance (bp) and number of haplotypes. C) Haplotype diversity, as measured by the number of alleles related to interval size D) Number of common haplotypes correlated to interval length.

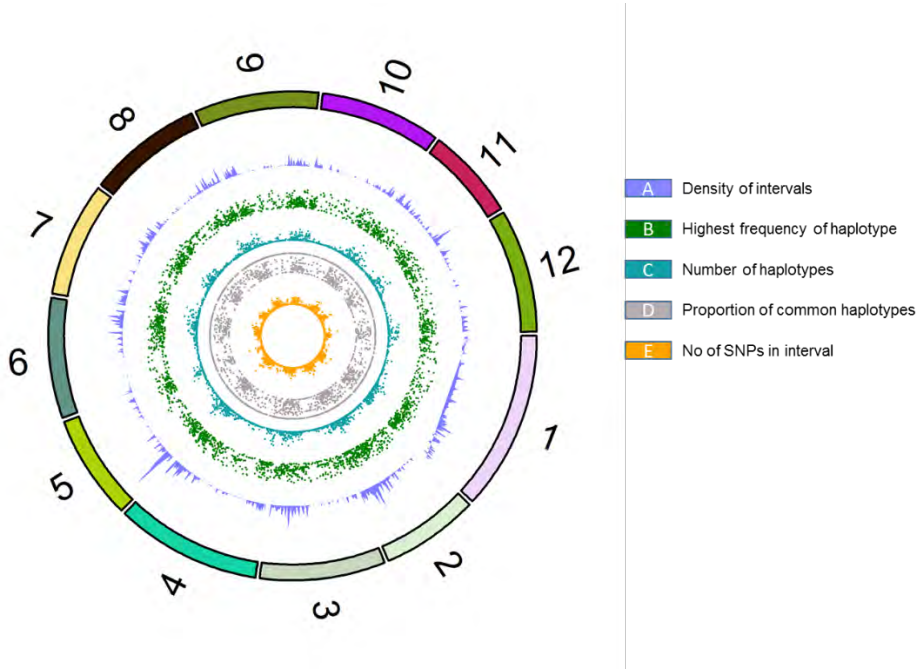


Figure 4. Genome-wide distribution of haplotype blocks. A) Genome-wide density of intervals used for haplotyping B) Frequency of the most common haplotype. C) Number of haplotypes per interval D) Proportion of (common) haplotypes with MAF > 0.05. E) Number of SNPs within an interval.

In total 20492 haplotypes were constructed across 1478 loci with on average 13.9 haplotypes per locus. The number of reconstructed haplotypes per locus correlated with the length of the interval (R^2 = 0.2035), as well as with the number of SNPs per interval (R^2 = 0.7451). The distribution of the number of haplotypes per locus is far from uniform. One tail of the distribution shows that 20% of all loci have no more three haplotypes, but this merely reflects the low average density of 2.09 SNP in those intervals. The other 20% tail of the distribution shows that these 295 loci have at least 8 up to 63 SNPs per interval (on average 18.6 SNPs) resulting in at least 24 up to 86 haplotypes (on average 36.0 haplotypes per interval), which merely reflects that besides true haplotypes also erroneous haplotypes accumulate both with interval size and number of SNPs (Figure 3C). In the absence of ground-truth haplotype information we cannot know the upper

limit, but the potato gene pool is both diverse and has accumulated mutational load. Therefore it is our impression that e.g. 24 haplotypes per locus is far from exceptional. Haplotypes with very low frequencies are not necessarily erroneous. Given the large number of wild potato species, used for introgression breeding for disease resistance, many haplotypes with an allele frequency below 1% can be expected. Among these low frequent haplotypes we expect new haplotypes due to introgression breeding and recombination events.

After a description of the number of haplotypes per locus and their allele frequencies, we studied the distribution of haplotype diversity across the genome. Figure 4 shows that pericentromeric regions are less diverse in the number of SNPs or haplotypes per interval than the gene rich distal regions. Furthermore, based on our input data from the SNP array, most intervals are located in gene rich regions for obvious reasons. Therefore we cannot conclude that genetic diversity is higher in gene rich regions. As visualized in Figure 4 haplotype diversity varies per locus and location across the genome.

PVY-resistance allele

To provide a more in-depth case study of the applicability of the haplotype inference method we tried to reproduce earlier work on a specific haplotype on chromosome *11*, introgressed from the wild potato species *S. stoloniferum* (CPC 2093), conferring resistance to two PVY strains, PVY^{O} and PVY^{NTN} (Van Eck et al. 2017). This introgression segment is present in a few varieties only. Phasing should identify a haplotype in the same varieties containing the introgression segment. In the previous study, this introgression segment was characterized using seven haplotype-specific SNPs (hs-SNPs), which were shared only between EOS, Y 66-13-636, and Festien, in a panel of 83 potato varieties. Subsequently genotypic information of these SNPs, in the panel of 537 potato varieties showed that 16 varieties should contain this introgression segment.

	PotVar0063974	PotVar0064012	PotVar0064036	PotVar0064037	PotVar0064044	PotVar0064045	PotVar0064046	PotVar0064080	PotVar0064083	PotVar0064140	PotVar0064141	PotVar0064142	PotVar0064152	PotVar0064155	PotVar0064177	PotVar0064182	PotVar0064192	PotVar0064200	solcap_snp_c2_13350	solcap_snp_c1_4296	PotVar0064345	PotVar0064391	PotVar0064400	PotVar0064415	PotVar0064426	PotVar0064470	PotVar0064472	PotVar0064473	PotVar0064474	PotVar0064475	PotVar0064502	solcap_snp_c2_13431	PotVar0064519	PotVar0064578
Y 66-13-636	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1
Y 66-13-636	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y 66-13-636	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
Y 66-13-636	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0

Figure 5. Haplotype composition of Y 66-13-636. The grey boxes indicate seven SNPs specific for the haplotype conferring PVY resistance. Allele one is the most rare allele.

We employed the haplotype inference method on all 27 SNPs spanning a 0.53 Mb interval flanked by PotVar0063974 to PotVar0064578 on chromosome *11* (Figure 5) and obtained the haplotype composition of all 537 varieties. For genotype Y 66-13-636 we reconstructed four different haplotypes. All seven haplotype-specific SNPs were linked to one haplotype, which also comprise the common SNP PotVar0064473, which is present in multiple haplotypes (Figure 5). For all other 19 SNPs present in the potato gene pool the reference allele was observed in the allele conferring resistance. We also studied the haplotypes in a pedigree structure with variety Cupido (resistant) and its parents W 72-22-496 (resistant) and Estima (susceptible). Figure 6 shows that the inferred haplotypes of Cupido are fully concordant with its parental varieties. To summarize, we could identify one haplotype for PVY resistance, among 36 inferred haplotypes, from an interval with 27 SNPs. All SNPs that specify this haplotype are identical by descent and trace back to the donor of the resistance, clone Y 66-13-636, except the SNP PotVar0064473. This common SNP has high allele frequency (47%) and was first observed in Pink Fir Apple (1850) and its polymorphism represents homoplasy caused by random nucleotide substitutions accumulating over time across potato species.

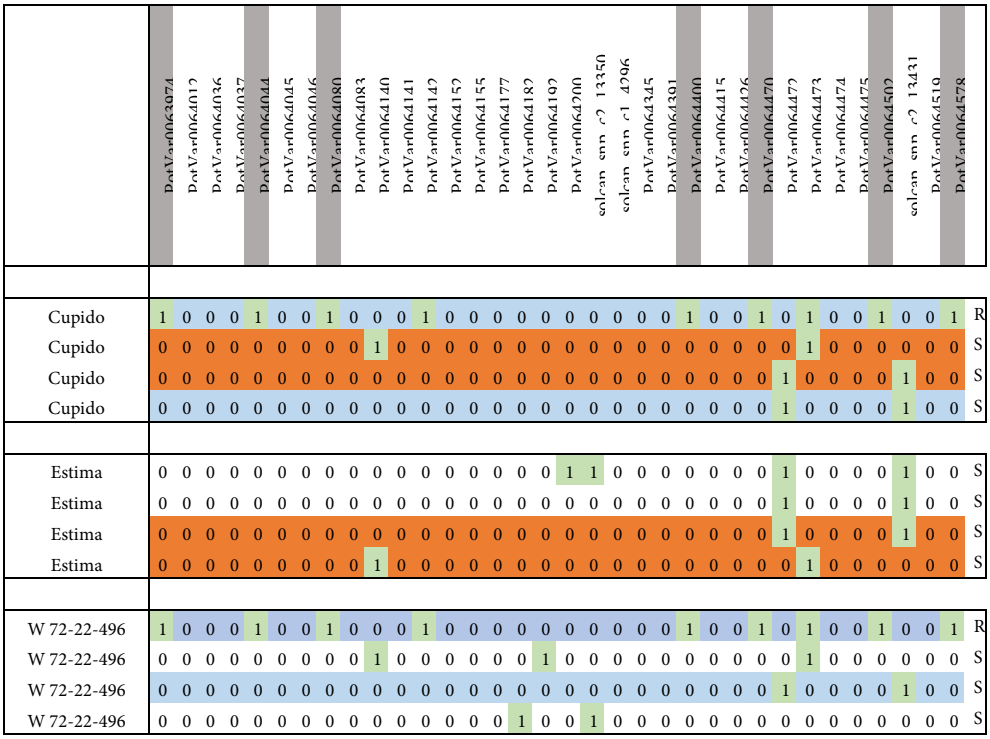


Figure 6. Validation of inferred haplotypes with variety Cupido (resistant) descending from W 72-22-496 (resistant mother) and Estima (susceptible father). The 'S' refers to susceptible alleles, and the 'R' referes to resistance alleles.

Discussion

We developed an approach to reconstruct haplotypes from un-phased SNP genotyping data from polyploids. In this study, we present a scalable approach which allows estimating haplotypes in intervals containing a large number of SNPs. Existing software, such as ShesisPlus (Shen et al. 2016), SATlotyper (Neigenfind et al. 2008) and polyHap (Su et al. 2008)., allow to compute haplotypes up to 10-20 SNP length, but require substantial computational resources to reconstruct longer haplotypes. A recent publication by Shen et al. (2016) estimated that both SATlotyper and polyHap require a unfeasible computation time to compute haplotypes for larger SNP numbers, and proposed the PL-EM algorithm for polyploids. Their approach implements a EM algorithm for haplotype inference, but uses a partition-ligation strategy to incrementally build larger haplotypes, avoiding the need to consider all possible haplotypes at a locus.

Here we explore the use of a stepwise approach, that uses partial phase information as estimated by the EM algorithm to reconstruct longer haplotypes. The first step computes all pairwise phasings that are possible for a set of SNPs. Increasing this number would result in *n-1* additional pairwise phasings to be computed, where *n* is the total number of markers in a segment. Subsequently, we use these pairwise phasings as input, and construct the full-length haplotypes SNP by SNP. This allows the scaling of our approach up to arbitrary SNP numbers.

Correct estimation of haplotypes depends on two factors: Firstly, the correct assignment of the allele dosage during genotype calling, and secondly the correct estimation of the linkage phase by applying the EM-algorithm. Although the error percentage for a single SNP is low for SNP array data, we clearly need to consider this. For instance, considering a segment of 20 SNPs, and a high genotyping accuracy of 0.99 for each SNP locus, we will observe that only 0.99^{20} = 81% of the varieties have completely accurate genotype calls. Hence, phasing errors will be inevitable. Likewise, phasing errors can occur due to ambivalence in assigning haplotype frequencies. The combination of dosage errors and phasing errors will greatly influence the ability to accurately reconstruct haplotypes.

To improve our haplotype reconstruction and to avoid errors, we repeat our haplotype reconstruction method multiple times. As each iteration chooses a different order in adding SNPs, this will result in slightly different haplotype reconstructions. After multiple iterations, the confidence of a certain haplotype reconstructions can be assessed and the solution can be discarded if a user-defined threshold is not reached. This was illustrated with the *StGWD1* amplicons, consisting of 78 SNPs. Without filtering approximately 98% of the reconstructed haplotypes are correctly estimated. With filtering, we could increase this to 100%, however in that case, approximately 10% of the genotypes were not inferred. The accuracy and missing call rate depends on the choice of the user-defined thresholds.

As any of the reconstructed haplotypes can contain a switch or base flips, validation of haplotypes is important. One approach to further validate haplotypes is the use of pedigree related samples (Figure 6), where identity-by-descent (IBD) allows to filter or to correct erroneous haplotypes. Filtering on frequency allows to reduce the number of phasing errors, as phasing errors are likely to occur randomly, resulting in low frequent

haplotypes, as exemplified by the reconstruction of alleles within the *StGWD1* amplicons, where erronous haplotypes have a low allele frequency. Likewise true haplotypes which occur at low frequency often escape detection, and cannot be differentiated from erroneous haplotypes.

Haplotypes in potato

Earlier studies in diploid organisms have shown that haplotype-based association studies have improved statistical power for QTL detection (Bakker et al. 2005, Schaid et al. 2008), or facilitate better design of haplotype-specific markers for use in marker-assisted breeding (MAB) (Buntjer et al. 2008).

In potato, it has been shown that haplotype blocks (as defined by historical recombination events) are large (> 0.5Mb < 2.5 Mb)in comparison with other plant or animal species, where LD is decayed after < 100 kb (Vos et al, 2017). One of the explanations for having such a long-range decay of LD is the limited number of (meiotic) recombination events that potato underwent since domestication. Based on LD decay estimates, of a simulated potato variety panel, it was previously suggested that a limited number of founder haplotypes (6-12) should suffice to generate the allelic diversity as present currently in the potato genepool (Vos et al. 2017). Here we reconstructed haplotypes across small intervals (< 30 kb), showing that a large number of haplotypes are reconstructed. Within these intervals, commonly, a restricted set of haplotypes with MAF > 0.05 cumulatively explains most of the haplotype diversity (92%), indicative of presence of only few haplotypes.

We used our haplotype reconstruction approach on 1478 intervals with 2-63 SNPs. Based on our reconstructions, we observe that the distribution of allele frequencies follows an exponential distribution. While this is to be expected in a random mating population, this is less likely within a population that arguably represents selected material. Genome-wide distribution of haplotype diversity shows that this estimate varies per locus and location (Figure 4).

In this study we reconstructed haplotypes which can be used to improve genetic studies in tetraploid diversity panels. So far many studies have only used single SNP markers to perform association mapping (Vos et al. 2015, Chapter 5, Rosyara et al. 2016). Each of these studies would benefit from the additional information that is present within the

haplotypes, primarily because of knowledge about haplotype-specificity of individual SNP markers.

Application of haplotype inference often lacks resolution to discover haplotypes that have low frequency (e.g. novel recombinants, introgressions), and differentiate between these and erroneous phasing results. Here we reconstructed haplotypes with 27 SNPs near the $Ny_{(o,n)sto}$ locus conferring resistance to *Potato Virus Y*. We identified a haplotype with frequency of 0.5% that is absent in susceptible varieties, but indicative of resistant varieties. By using pedigree relations within our variety panel we could demonstrate correct haplotype reconstruction. This indicates that the use of pedigree relations allows to verify the presence and correct reconstruction of rare alleles.

Further development of haplotype inference methodology is required.

Despite our focus on tetraploid potato we developed a method that can be applied irrespective of the ploidy level. Genotypic data of other species with a different ploidy level can be used as well as input for our algorithm. Opposed to SATlotyper and Polyhap, of which application allows to reconstruct haplotypes only for limited SNP numbers, our new algorithm can reconstruct haplotypes over arbitrary length. So far we only used a genotyping dataset of 14K markers, and were able to processed intervals containing up to 78 SNPs. A downside of our algorithm is that it is highly dependent on the quality of underlying dosage calls. This limits the application to dosage calls originating from next-generation sequencing data with inadequate read depth or allele bias.

To make this algorithm suitable for the use of error-prone genotyping data such as sequencing-based dosage calls, an improvement could be achieved by modelling ambivalence in dosage assignment in the joining step. Last but not least, in many diploid studies, it is suggested that the use of IBD relations greatly improves haplotype accuracy (Garg et al. 2016; Motazedi et al. 2017). This will provide information about the inheritance or expected segregation patterns of alleles. For instance, haplotype reconstruction accuracy could be improved by only allowing haplotype solutions that are in agreement with the pedigree structure. This likely would result in greater power to determine the correctness of low-frequent alleles. In our case, we likely could implement this by penalizing haplotype solutions that are not in agreement to genetic relationships present within the whole panel.

Concluding remarks

So far a scalable approach for polyploid haplotype inference was not available, limiting application of haplotypes in polyploid genetic studies. This study demonstrates that accurate haplotype inference can be achieved by using un-phased genotype information from polyploid species. Our approach uses a simple EM-based estimation step to estimate linkage phases, coupled with a stepwise joining algorithm that allows computing haplotypes of arbitrary lengths, in a stepwise at SNP by SNP. Our results indicate that haplotype inference is useful to obtain more information about the haplotype composition of regions for which only SNP data is available. In potato we observe few common alleles, explaining a large proportion of the allele diveristy, and many rare alleles. To distinguish between phasing errors and correctly estimated low-frequent rare alleles pedigree data can be used. Previously it has been suggested that for application of marker-assisted selection in potato breeding markers should be haplotype-specificity. The application of this tool will therefore provide a valuable contribution for successful application of maker-assisted selection in potato.

Acknowledgements

JHW is supported by a grant of the Dutch Science Organisation NWO (project 831.14.002). The Dutch Technology Foundation (STW grant WPB-7926) financed the development of the SolSTW array and SNP-data production. We thank Michiel Klaassen for critical feedback on this manuscript.

Additional files

File S1 Dosage calls of *StGWD1* amplicons.

File S2 Ground-truth haplotypes of *StGWD1* amplicons.

File S3 Reconstructed haplotypes of *StGWD1* amplicons.

File S4 Overview 1479 intervals.

Chapter 5

Haplotype-based genetic analysis identifies relevant alleles for agronomical traits in potato

Johan H. Willemsen, Yiyuan Ding, Peter G. Vos, Richard G.F. Visser, Herman J. van Eck.

Abstract

A potato diversity panel composed of 537 tetraploid potato varieties was analysed for significant allele-phenotype associations using inferred haplotypes. So far genetic analysis in tetraploid potato has been performed using single-marker association analysis, because the underlying haplotype structure was unknown. Here we use haplotype information in autotetraploid potato to revisit previously observed QTLs for four representative traits.

The detected QTLs not only confirmed the outcomes of earlier studies (i.e. for plant maturity, flesh color and tuber shape), but also led to the identification of novel putative QTLs for plant maturity and flesh color. Based on the haplotype structure at these QTLs we link phenotypic variation to specific alleles of known candidate genes. In addition, on the basis of our results, we propose that knowledge of allele composition, and haplotype-specificity of single SNPs is crucial to QTL discovery and to correctly understand and interpret association analysis. The advantages and disadvantages of haplotype-based genetic analysis are discussed, and we suggest to combine both analyses in order to successfully identify alleles involved in agronomical traits, as a prerequisite for successful implementation of marker assisted breeding in a polyploid species.

Keywords: association mapping, genetic analysis, haplotypes, alleles, haplotype-specificity, potato

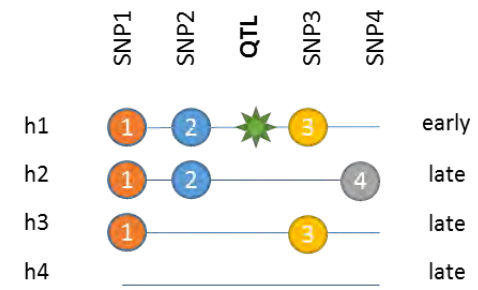
Introduction

Genome-wide association studies (GWAS) have been performed successfully in potato, identifying loci involved in monogenic traits such as plant maturity (Kloosterman et al. 2013), tuber shape (Chapter 2, Sharma et al. 2018), yellow flesh color (Vos et al. 2017; Rosyara et al. 2016) and disease resistances such PVY resistance (Van Eck et al. 2017). Likewise, also for more complex polygenic traits such as glycoalkaloid content (Vos et al. 2017) and other agronomical traits multiple QTLs were discovered (D'hoop et al. 2014; Rosyara et al. 2016; Malosetti et al. 2007; Urbany et al. 2011). Each of these association mapping experiments used a variety panel of varieties, and identified significant marker-trait associations. These markers can subsequently be applied in marker-assisted selection to accelerate breeding progress. However, the application and interpretation of these marker-trait associations are challenging as often the underlying haplotype structure is unknown. The individual SNP marker allele(s) may be specific for one haplotype (e.g. a so-called haplotype-specific SNP (hs-SNP)), or could be present across multiple haplotypes, which collectively might result in a series of multiple alleles. In Figure 1 a scenario is visualised where individual SNP markers are present in multiple alleles, with as consequence, diminishing the predictive value of these SNPs for presence of a specific allele. Each of these alleles can have its own contribution to the trait value (Uitdewilligen et al. 2012). Some examples of these allelic series were previously observed in potato (van Eck et al. 1994; Uitdewilligen et al. 2012; Schreiber et al. 2013), suggesting that all haplotypes at a locus should be considered jointly during association mapping. In other crops, examples are found for presence of allelic series such as in the diploid apple (Di Guardo et al. 2017) and hexaploid chrysanthemum (van Geest et al. 2017).

In an association mapping study, several factors influence the statistical power to detect QTLs. Firstly, genetic variation at different loci could lead to similar phenotypes (genetic heterogeneity), and secondly, distinct alleles at the same locus could lead to similar phenotypes (Bergelson and Roux, 2013). This so-called allelic heterogeneity will likely influence the power to detect QTLs, with individual SNP markers, because the underlying multiple alleles, are not modelled during QTL detection. Thirdly the power of QTL detection is for the most part determined by the genetic architecture of the trait (Ingvarsson & Street, 2011). Some traits are largely monogenic, and controlled by major

effect QTLs, allowing rapid detection of QTLs, whereas others are controlled by a large number of minor effect QTL, requiring large genotype panels to achieve enough statistical power for QTL discovery.

As crops often display extensive population structure, the interpretability of association studies, is hindered by false positive associations, due to correlation of a trait with population structure, and to counteract this a correction for population structure is required (Kang et al. 2008, Malloseti et al. 2005).



A

	S1	S2	S3	S4
H1	0.33	0.5	0.5	0.33
H2	0.33	0.5	0.5	1.0
H3	0.33	0.5	0.5	0.33
H4	1.0	0.5	0.5	0.33

B

Figure 1. Haplotypes have more discriminatory power to detect alleles than individual SNPs. A) H1 is the causal allele. In this case the QTL (star) is not genotyped, but SNP1-4 are. B) Discriminative power of individual SNPs for haplotypes. Only SNP1 and SNP4 are haplotype-specific. To discriminate between the causal H1 allele and other non-functional alleles a combination of SNPs is needed.

Reconstructing haplotypes

The successful application of haplotypes for QTL discovery in polyploids depends on having accurate haplotype information. Conventional genotyping platforms provide genotypic information of individual SNPs, but do not allow to obtain linkage between SNP-alleles. So far most studies in potato that apply allele-phenotype association methods have focused on allele identification from PCR amplicons (Simko et al. 2004; Schreiber et al. 2013; Uitdewilligen et al. 2012), which subsequently are used to detect significant marker trait associations. However, in a previous study we introduced an approach for haplotype inference for polyploids which allows to compute haplotypes for multiple adjacent bi-allelic SNP markers (Chapter 3). This approach relies on an Expectation-Maximization (EM)-based approach for determining the linkage phase between alleles of adjacent SNPs, considering the haplotype frequency within all

varieties, followed by a SNP-by-SNP extension of each haplotype within a single variety. This haplotype inference method is sensitive to genotyping error. Dosage errors will result in one or multiple erroneous phased haplotypes. With increasing haplotype length or SNP numbers, the probability of a genotyping error increases, as well as the amount of erroneous haplotypes. Therefore it is expected that haplotype-based GWAS is affected by the uncertainty of haplotype reconstruction (Stram and Seshan, 2012).

This study

In this paper we explore the comparison of a GWAS with single SNPs and a GWAS using haplotypes as substitute for these single markers, to test if haplotypes provide greater discriminatory power to detect QTLs. Knowledge of haplotype-specificity of each SNP within a haplotype allows us to investigate the question if haplotype-specificity is important for adequate marker-trait association within this genotype panel. For all markers within a QTL region we can subsequently determine the suitability of these markers to tag the underlying alleles. We subsequently discuss the insights gained and propose a balanced use of haplotype-based analysis in combination with single marker GWAS to improve genetic analysis in polyploids. Here we show that identification of haplotype-specificity of SNP-alleles will be instrumental for the application of marker-assisted selection.

Methods

Genotypic and haplotype datasets

A panel of 537 potato varieties was genotyped using with the SOL-STW Infinium SNP array (Vos et al. 2015). In short, genotyping was performed using fitTetra (Voorrips and Maliepaard, 2008), using raw signal intensities as input. These genotype calls were used to obtain a total of 15K informative SNP markers. We reconstructed haplotypes with the approach described Chapter 3. Haplotypes were reconstructed over a sliding window comprising 10 SNPs. Each of these segments of 10 SNPs is considered a separate haplotype block and was used in association mapping.

We used largely monogenic traits such as flesh color, tuber shape and plant maturity (daylight-dependent tuberization) as model traits to validate our approach. In addition we evaluated potato tuber uniformity as an example of a polygenic trait. Phenotypes

were scored as described in 'd Hoop et al. (2008, 2014), over multiple years and locations, and for each trait, the best linear unbiased estimates (BLUEs) were estimated. Uniformity was estimated by sorting tubers on sizes, and measuring average sizes in classes 0-30, 30-40, 40-50, 50-60, 60-70 and 70+ mm. The mean tuber size of each class was multiplied by the number of tubers present in each class, and all values were summed, resulting in a single value for uniformity. These values are divided by the total number of tubers per variety.

Association mapping

Single marker GWAS was performed with GWASpoly (Rosyara et al. 2016), with correction for population structure. For haplotype-based association we modified GWASpoly to allow the inclusion of haplotypes as predictor variables. Hence we modelled quantitative trait variation by considering the contributions of each of the possible haplotypes present at a locus. At any locus the contributions of the i^{th} allele can be represented by a dummy variable X_i, where i E (1, 2 … k). If $X_i = 0$ then the haplotype is not present in an individual, if Xi is larger than 1, this implies that the allele is present with a dosage more . The model that was used can be described as:

$$y_{ij} = \mu + X_i + \underline{K} + \varepsilon_{ij} \quad (i = 1 \ldots k\,; j = 1..n)$$

Where y_j is the phenotype value of an individual, μ is the overall mean, K represents population structure determined by the kinship matrix, and ε_{ij} is the residual error term (genotypic association). We subsequently used a general F-test to test if any of the haplotypes have an effect significantly different from zero. To estimate which haplotype has an influence on trait variation, the regression model was performed by using a single haplotype in the regression model (allelic association). To determine how much of the phenotypic variation is explained, we also calculated the squared correlation of each SNP or haplotype to each phenotype.

Haplotype-based association mapping

We conducted GWAS analyses using the panel of 537 genotypes for all traits with both a naïve GWAS analysis, and GWAS with correction for population structure. We explored if haplotype-based GWAS does identify more QTLs than single SNP GWAS. For each trait this results in three association mapping models that were performed in a

single haplotype block of 10 SNPs (Table 2). Firstly, regression analysis of 15K single SNP markers. Secondly multiple regression of all haplotypes at 15K loci. Thirdly allelic regression of a single haplotype on phenotypes for all haplotypes present at all 15K haplotype blocks.

Description	Hypothesis	Statistical tests per block
Single marker regression	H0: no effect H1: effect	10
Multiple haplotype regression	H0: no effect H1: one or more haplotypes has significant effect	1
Allelic regression	H0: no effect H1: haplotype has effect	# haplotypes

Table 1. Three models for association mapping were used and evaluated

We defined a putative QTL if any of the above specified analysis results in a clear QTL peak with multiple markers/haplotypes at a threshold of $-\log_{10}(p)$ of 4.0, which appeared adequate for control of false-positive associations. For monogenic traits with a major-effect QTLs, a more detailed analysis was performed by comparing the most significant haplotype blocks to the marker trait associations of SNPs present within these QTLs.

Demographic history of alleles

Previously each SNP was dated by using the year of market release of the oldest variety that first displayed the presence of this minor SNP allele (Vos et al. 2015). Here we apply the same procedure to give a date to haplotypes. The year of introduction was based on year of market release as found in the potato pedigree database (Van Berloo et al. 2007; https://www.plantbreeding.wur.nl/PotatoPedigree/). This procedure allows to determine which haplotypes are recently introgressed from wild species, and probably display identity-by-descent (IBD) (post-1945) and haplotypes that represent old standing genetic variation (pre-1945).

Results

Analysis of phenotypic data

Substantial phenotypic variation was observed for all traits, which for maturity has a broad-sense heritability of 0.85, for tuber shape 0.74 and for flesh color 0.63 (D'hoop et al. 2014). Potato tuber uniformity has a heritability of 0.46 and is significantly correlated with 'year of release' of a variety. We explored correlations between traits and three structure groups (rest, starch, processing) as defined by D'hoop et al. (2011) and Vos et al. (2016). In view of previous association analysis, presented in Chapter 2, we conclude that tuber shape is correlated with structure groups, as varieties used by the starch industry are significantly more round (6.84 ± 0.14), than in other structure groups Agria (5.01 ± 0.18) and Rest (4.98 ± 0.08). The same pattern is observed for maturity, where the 'Starch' group generally has a longer growing season (4.84 ± 0.15), than varieties that belong to Agria (6.29 ± .10) and Rest group (6.29 ± 0.05). For flesh color similar values are found for two groups, 'Rest' and 'Starch', and only the Agria group has more yellow flesh color (6.71 ± 0.05). Potato uniformity has lower scores for uniformity (5.91 ± 0.08) in 'Starch' potato varieties, compared to varieties that belong to 'Agria' (6.33 ± 0.037) and 'Rest' group (6.48 ± 0.52), which largely can be explained by the lack of selection criteria for uniformity within starch breeding programmes.

Traits	Complete (n=537)	Starch (n=59)	Rest (n=407)	Agria (n=71)
Shape	5.93 ± 0.07	6.84 ± 0.14	4.98 ± 0.08	5.01 ± 0.18
Maturity	6.14 ± 0.05	4.84 ± 0.15	6.29 ± 0.05	6.29 ± 0.10
Flesh Colour	6.08 ± 0.04	5.67 ± 0.1	6.71 ± 0.05	6.02 ± 0.08
Uniformity	6.30 ± 0.02	6.41 ± 0.07	6.88 ± 0.45	6.70 ± 0.03

Table 2. Averages of traits for the complete panel and sub-populations

Haplotype data analysis

As input for haplotype-based association mapping we used haplotype blocks over a sliding window, comprising 10-SNPs, resulting in 14409 blocks that partially overlap. The justification of choosing 10 SNPs as window length is made as results from earlier analyses show that haplotype reconstruction accuracy is influenced by number of included single SNP markers, and computational speed. Each block was filtered on MAF < 0.01 and haplotypes with low confidence were discarded. Within these blocks an average of 8.5 unique haplotypes was observed, from which 54% (MAF > 0.05) can be considered haplotypes that are common haplotypes and 46% are alleles that can be considered rare (MAF < 0.05).

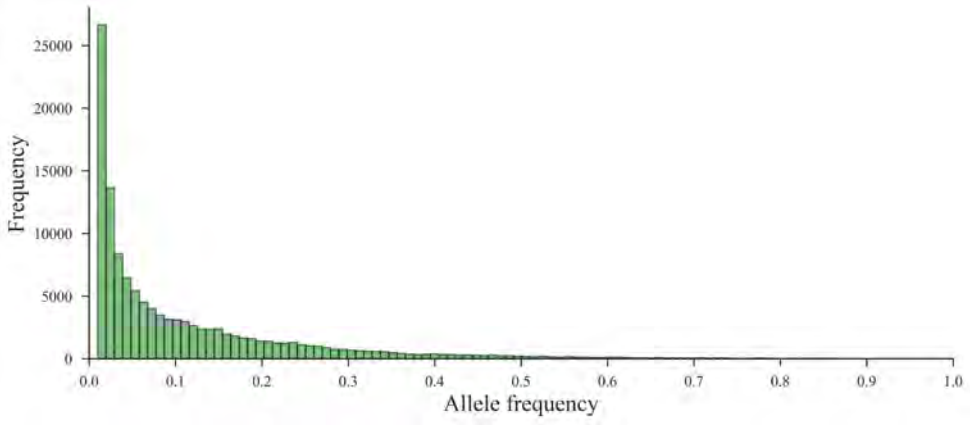


Figure 2. Allele frequency distribution of reconstructed alleles (N=121K). Most alleles have an allele frequency < 0.05 (54%), whereas a smaller proportion has a allele frequency > 0.05 (46%).

Association mapping

The haplotype-based analysis was performed for all traits. Based on these analysis we conclude that most QTLs were detected in both single marker analysis as haplotype-based association mapping. In addition to the location of the QTL we identified which haplotype(s) have a significant effect on the phenotype. Here we describe these results in more detail. We restrict the detailed analysis to the findings for traits with a simple genetic architecture (plant maturity, tuber shape, flesh color), and provide an overview of the findings for uniformity. As potato exhibits strong population structure, and diagnostic Q-Q plots (File S1) show substantial *p*-value inflation in absence of kinship-correction, we only report the kinship-corrected results.

Flesh color

A well-characterized trait in potato is (yellow) flesh color, which is known to be largely regulated through allelic variation within the *StCHY2* gene (Wolters et al. 2010). Single marker association without kinship correction, identified a very strong QTL on chromosome 3 at two markers: PotVar0070260 and PotVar0120627 with $-\log_{10}(p)$ of 17.9 and 19.86 respectively. The location of these SNPs is 43921937 and 48550473 on potato chromosome 3, respectively. With multiple haplotype regression, the most significant association was found at similar position, where block_4027 (St4.03ch03:43631694-43922172), approximately 200 kb upstream of PotVar0070260 had the strongest association with flesh color with $-\log10(p) = 26.5$. This haplotype block contains 10 haplotypes, and the presence of all 10 haplotypes of block 4027 explains 48% of the phenotypic variation. Strikingly, the application of allelic regression showed that the strongest allele-phenotype association is observed for a single haplotype within block_4246 (St4.03ch03:48792194-49236778) with $-\log10(p) = 25.1$, explaining 37.1% of the phenotypic variation. This block is located at considerable distance downstream of block_4027 (+ 4.8 Mb). The presence of the StCHY2.1 haplotype of block_4246 allows to predict yellow and white flesh color (Figure 4), although for varieties that have increased dosages of this allele (>2) a darker yellow flesh color is observed, suggesting incomplete dominance. Within our panel we could not observe a significant effect of the 5 remaining haplotypes.

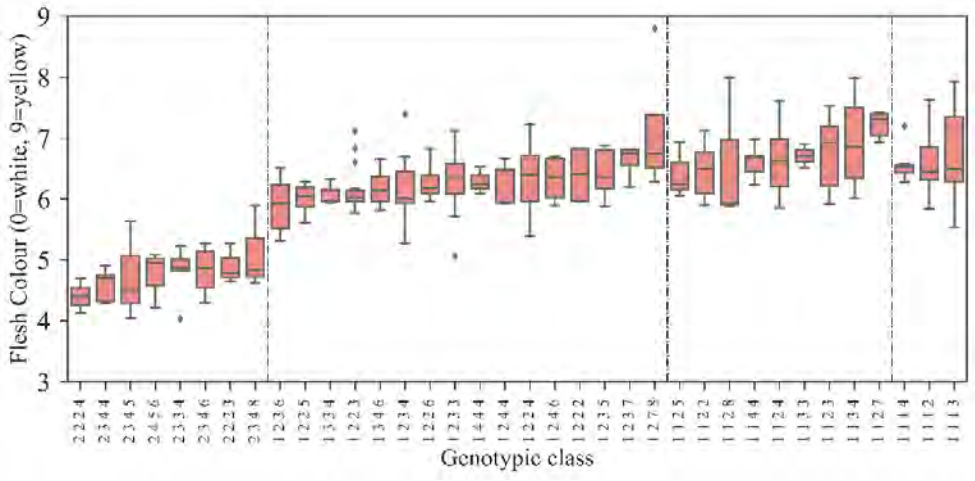


Figure 4. Distribution of phenotypic scores for flesh color related to haplotype composition at the StCHY2 gene. Each boxplot represents a genotypic class. Only genotypic classes with more than two observations are visualized.

In contrast, single marker association at PotVar0070260 and PotVar0120627 explains 52% and 51% respectively of the variation in flesh color. Judged from the haplotype structure of these blocks, PotVar0070260 and PotVar0120627 represent haplotype-specific SNPs that tag the *StCHY2.1* allele. Inspection also revealed that blocks in close vicinity of PotVar0070260 all show a single haplotype that is significantly associated with flesh color, with similar frequency. Although decay of linkage disequilibrium between overlapping haplotype blocks is not known, this likely represents the same allele. As already determined by Vos et al. (2015) PotVar0070260 was first seen in variety 'Yam'. From figure 5B we observe that allele (*StCHY2.1*) is present predominantly in more recent introduced varieties (> 1945) which have a higher dosage of this allele, compared to old varieties (< 1945).

Trait	Position	-log10(*p*)	Name	MAF	-log10(p)	Effect of allelic association
Flesh Color	ST4.03ch03	16.59	*StCHY1.1*	*0.29*	19.48	0.50
	48791760-48793081		*StCHY1.2*	*0.21*	0.01	0.00
	block_4027		*StCHY1.3*	*0.13*	0.67	-0.08
			StCHY1.4	*0.13*	1.42	-0.13
			StCHY1.5	*0.05*	2.41	-0.26
			StCHY1.6	*0.04*	0.33	-0.08
			StCHY1.7	*0.03*	4.79	0.47
			StCHY1.8	*0.02*	1.04	-0.22
			StCHY1.9	*0.02*	0.09	-0.04
			StCHY1.10	*0.02*	1.09	0.34
Flesh Color	ST4.03ch08	7.78	*CHR8.1*	0.28	7.94	-0.41
	47377310-47376238		*CHR8.2*	0.20	2.18	0.24
	block_10355		*CHR8.3*	0.19	0.88	0.09
			CHR8.4	0.09	1.08	0.27
			CHR8.5	0.09	0.27	0.17
			CHR8.6	0.04	0.02	-0.16
			CHR8.7	0.02	1.21	0.46
			CHR8.8	0.02	1.92	0.60
			CHR8.9	0.01	1.88	0.91
Tuber shape	ST4.03ch10	16.00	*Ro1*	0.70	15.38	-0.94
	48717669-48593621		*Ro2*	0.20	2.63	0.51
	block_12201		*Ro3*	0.03	4.00	1.24
			Ro4	0.03	1.98	1.47
Tuber shape	ST4.03ch02	3.84	*Ro2.1*	0.25	0.26	-0.14
	29093511-28751201		*Ro2.2*	0.21	2.64	-0.32
	block_2353		*Ro2.3*	0.17	0.07	-0.07
			Ro2.4	0.11	0.30	0.32
			Ro2.5	0.06	0.33	0.22
			Ro2.6	0.05	0.25	0.21
			Ro2.7	0.03	1.49	-0.77
			Ro2.8	0.02	3.01	0.71
			Ro2.9	0.02	0.24	0.85
			Ro2.10	0.01	0.66	0.35
			Ro2.11	0.01	2.43	1.02
Plant maturity	ST4.03ch05	20.25	*StCDF1.2*	0.27	20.70	0.81
	4489590-4488075		*StCDF1.1.11*	0.23	0.60	-0.17
	block_6650		*StCDF1.1.12*	0.19	3.15	-0.26
			StCDF1.1.13	0.09	1.19	0.00
			StCDF1.1.14	0.07	2.35	-0.45
			StCDF1.1.15	0.06	3.29	-0.51
			StCDF1.1.16	0.02	0.30	-0.04
			StCDF1.1.17	0.02	0.91	0.63
			StCDF1.1.18	0.02	0.44	0.04
Plant maturity	ST4.03ch03	4.99	*CHR3.1*	0.27	1.06	-0.09
	43921286-43326982		*CHR3.2*	0.25	5.18	0.49
	block_4021		*CHR3.3*	0.12	0.13	-0.14
			CHR3.4	0.11	0.01	-0.10
			CHR3.5	0.05	1.17	-0.39
			CHR3.6	0.03	3.30	-0.94
			CHR3.7	0.03	1.11	-0.21
			CHR3.8	0.01	0.00	0.32

Table 3. Haplotype-based QTLs detected for flesh color, tuber shape and plant maturity, only kinship-corrected results are shown. These blocks were selected because they display the strongest association in haplotype-based regression.

Apart from the major effect QTL co-localizing with *StCHY2,* a minor effect QTL was observed on chromosome 8. Significant marker-trait associations were observed for a region of 3 Mb located at 45 to 48Mb, with the most significant association at PotVar0103331 (CHR8:47376966) with -log10(*p*) of 5.4 and explaining 19% of the phenotypic variation. The multiple haplotype regression identified the strongest association (-log10(*p*) = 7.78) at block_10355, located at interval CHR8: 47.376.238-47377310. The combination of all haplotypes at this block explains 22% of flesh color variation. Within this haplotype block the allelic regression identified a single allele with frequency of 0.17, which on its own explains 17% of the phenotypic variation. The increase in dosage of this allele results in more white flesh color (Figure 5A), although the peak significant allele-phenotype association was found in block_10357 with -log10(*p*) of 11.7 and explained variance of 19%. Like with the *StCHY2.1* allele, this haplotype is present in higher dosages in recent varieties (Figure 5C). The joint presence of the *StCHY2.1* allele, and this allele at block_10357, explain together 48% of the phenotypic variation, which given the individual contributions of both alleles (37.1% and 17%) is increased (File S2, Figure 5).

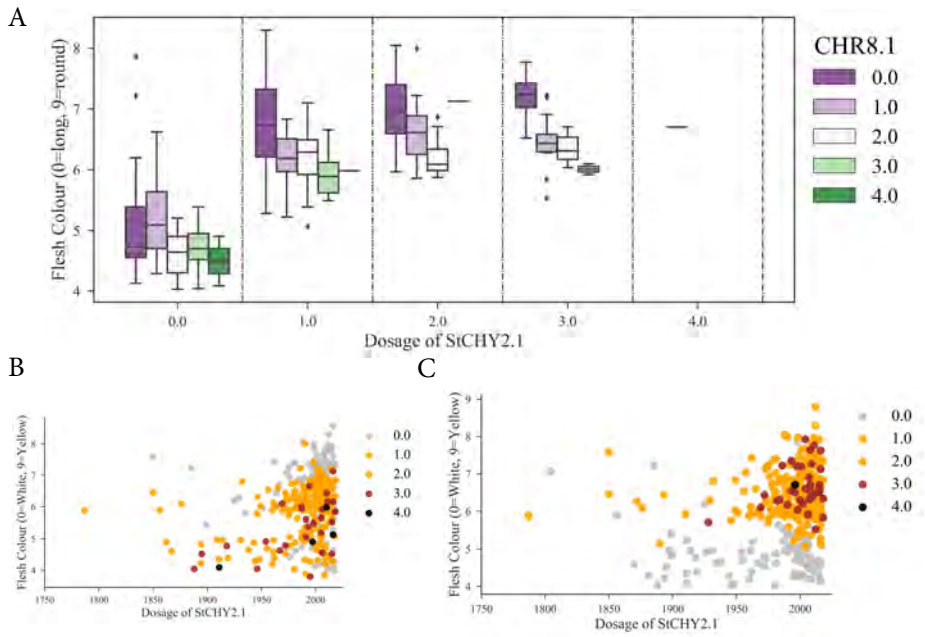


Figure 5. Distribution of trait values of Flesh color for all genotypic classes observed for the joint presence of the CHR3 QTL (StCHY2.1) and chromosome 8 QTL. B) Phenotypic distribution of significant allele-trait association at StCHY2.1. C) Phenotypic distribution of significant allele-trait association at CHR8.1 QTL.

Potato tuber shape

For tuber shape two QTLs were detected: a major QTL on chromosome 10 and a minor QTL on chromosome 2, in line with previous studies (Prashar et al. 2014, Chapter 2). Both QTLs have been found using haplotype-based analysis and single marker analysis (Figure 11). With single SNP association mapping we observed the most significant association for PotVar0111687 with $\log_{10}(p)$ of 18.7 in the kinship-corrected analysis. The haplotype-association did identify this QTL in block_12201, where four haplotypes were reconstructed. This block also contained PotVar0111687.

In this block four haplotypes (*Ro1*, *Ro2*, *Ro3*, *Ro4)*, were found with frequency 0.71, 0.20, 0.026 and .025 respectively. Only *Ro1* was found to have a negative effect on tuber shape (-0.89), implying that tuber varieties that contain this haplotype are significantly more elongated. For the three remaining haplotypes we determined the phenotypic effect, and presence of these other haplotypes result in rounder tuber shape, although the magnitude of their effects is different (Ro_2 + 0.51, Ro_3 + 1.24, Ro_4 + 1.47). The combination of these alleles explains 30% of the phenotypic variation. Allele *Ro1*, conferring long tuber shape, explains 25% of the variation, whereas individual alleles Ro2-4 explain 5%, 5%, 7% of the variation respectively. From Figure 6A we observe that allele (Ro_1) conferring long tuber shape, exhibits a slight recessive effect, where quadruplex scores of this allele lead to substantial longer tuber shape. In addition, the first two alleles (Ro_1, Ro_2) represent old variation, whereas the third and fourth allele (Ro_3, Ro_4) represent recently introgressed haplotypes (Figure 6B).

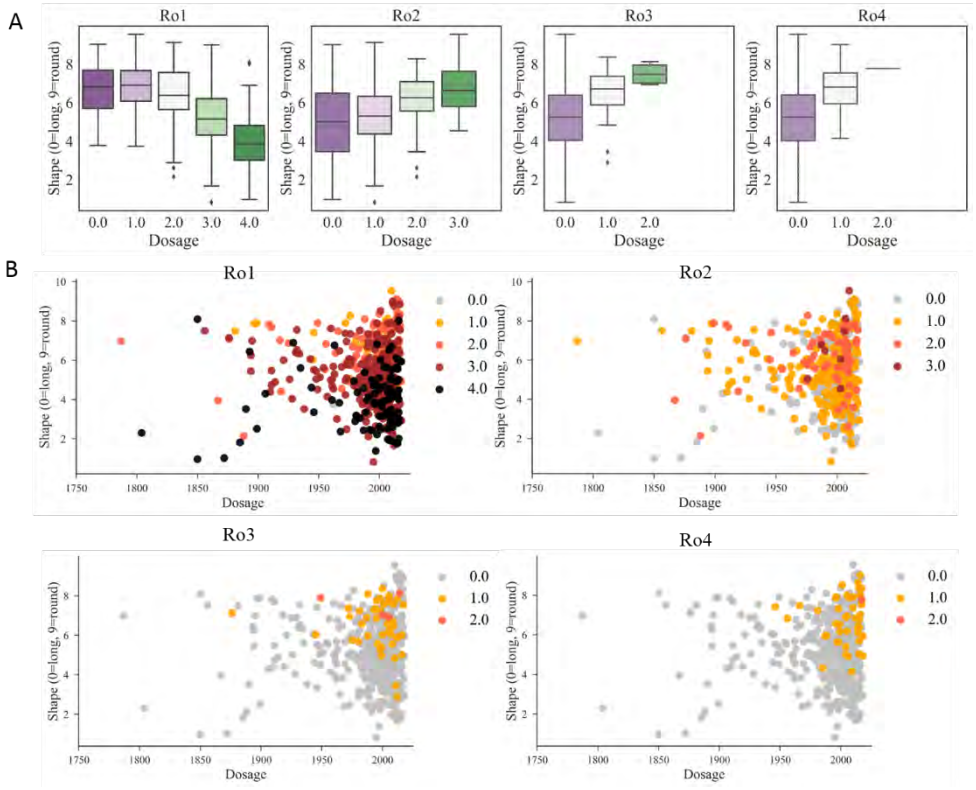


Figure 6. Haplotype analysis for chromosome 10 QTL for tuber shape (0=long and 9=round): A) Dosage effect of haplotype Ro1-4 on tuber shape, B) The first two alleles (Ro1, Ro2) represent old variation, whereas the third and fourth allele (Ro3, Ro4) represent recently introgressed haplotypes.

On chromosome *2* Prashar et al. (2014) andpreviously identified a minor QTL for tuber shape was identified. The most significant association was found within an interval between 27.60 and 28.04 Mb using single marker GWAS, at PotVar0123847 with $-\log_{10}(p) = 3.58$ and solcap_snp_c1_11556 with $-\log_{10}(p) = 3.99$. With haplotype-based analysis we identified an association at block_2353 with $-\log_{10}(p) = 3.84$ (Table 3). Surprisingly the allelic association did not result in any individual haplotypes associated with tuber shape.

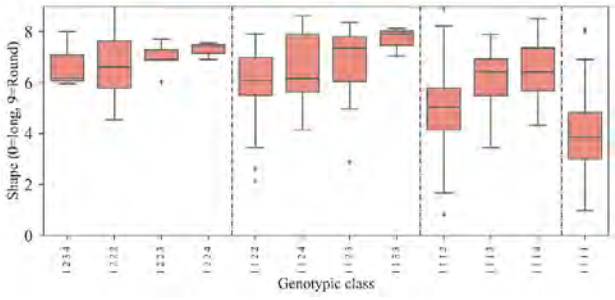


Figure 7. Distribution of phenotypic scores for tuber shape (0=long, 9=round) related to haplotype composition at the *Ro* locus. Each boxplot represents a genotypic class. Only genotypic classes with more than two observations are visualized.

Plant maturity

Previously it was postulated that for the major-effect QTL for early maturity, on chromosome 5, two alleles of the *StCDF1* locus contribute to the differences in plant maturity (Kloosterman et al. 2013). Here the most significant marker associated with plant maturity is PotVar0079081, which is 50kb upstream the *StCDF1* gene. In the kinship-corrected analysis single marker analysis a -$\log_{10}(p)$ of 23.38 is found for this SNP. This marker is present in block_6650 which was found to have the strongest association while using multiple regression of haplotypes. The combination of all haplotypes explains 31% of the phenotypic variation. From the 9 haplotypes in this block, only one (*StCDF1.*A*)* has a strong effect on earliness and explains 26% of the phenotypic variation (Figure 8A).

In addition to the major effect QTL localized at the *StCDF1* gene, haplotype-based GWAS identified a second QTL, located on chromosome 3 within block block_4021. This QTL was not observed in the association mapping results using single SNP markers. The multiple regression resulted in an association with $-\log_{10}(p)$ of 4.98, explaining 18% of the phenotypic variation. Within this block 8 haplotypes are reconstructed, from which only a single allele has an effect on maturity with $-\log_{10}(p)$ = 5.18, and explains 13% of the phenotypic variation. This single allele has a frequency of 0.25 and has a positive effect on earliness (Figure 9AC). Surprisingly none of the SNPs present within this interval tag this allele.

We subsequently used the two significant alleles of both QTLs, respectively *StCDF1.1* and CHR3.1 to express plant maturity as a combination of these haplotypes (Figure 10).

The combination of both QTLs explains 33% of the phenotypic variance. In addition a significant interaction was found of these alleles (*p-value* < 0.0043) (File S2).

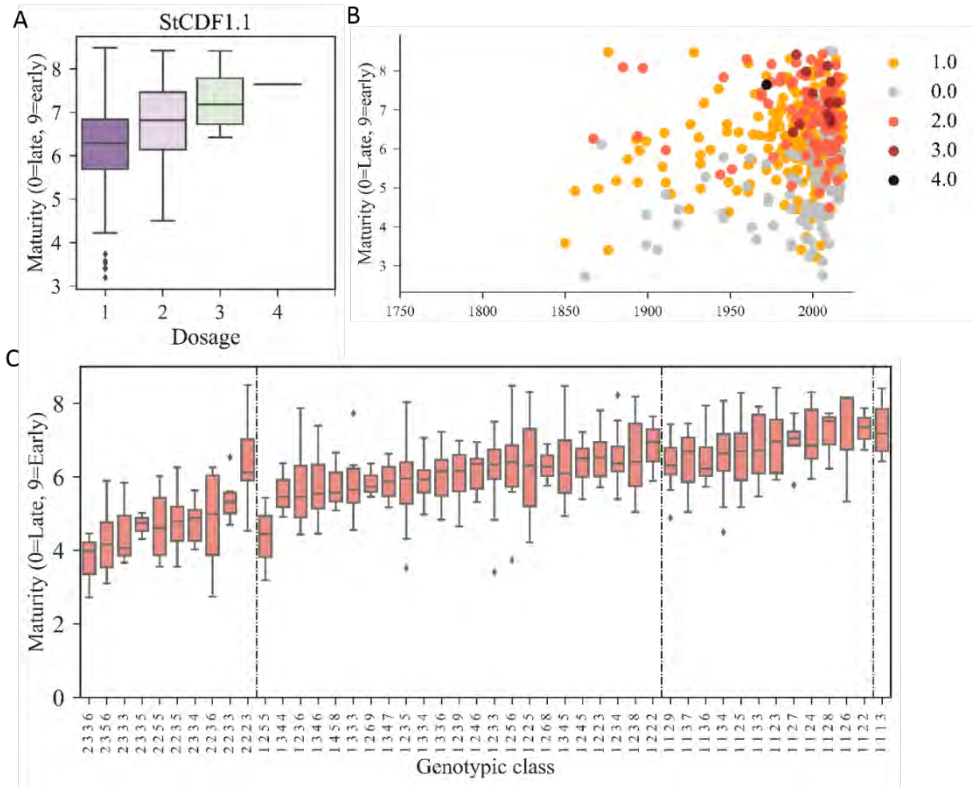


Figure 8. Haplotype analysis for chromosome 5 QTL for plant maturity (3 being late 8 early): A) Dosage effect of haplotype *StCDF1.A* on plant maturity, B) Haplotype *StCDF1.A* is present in both old and newly introduced varieties. C) Average phenotype scores per class. The '1' represents the allele that is responsible for earliness. Absence of this allele leads to late tuberization.

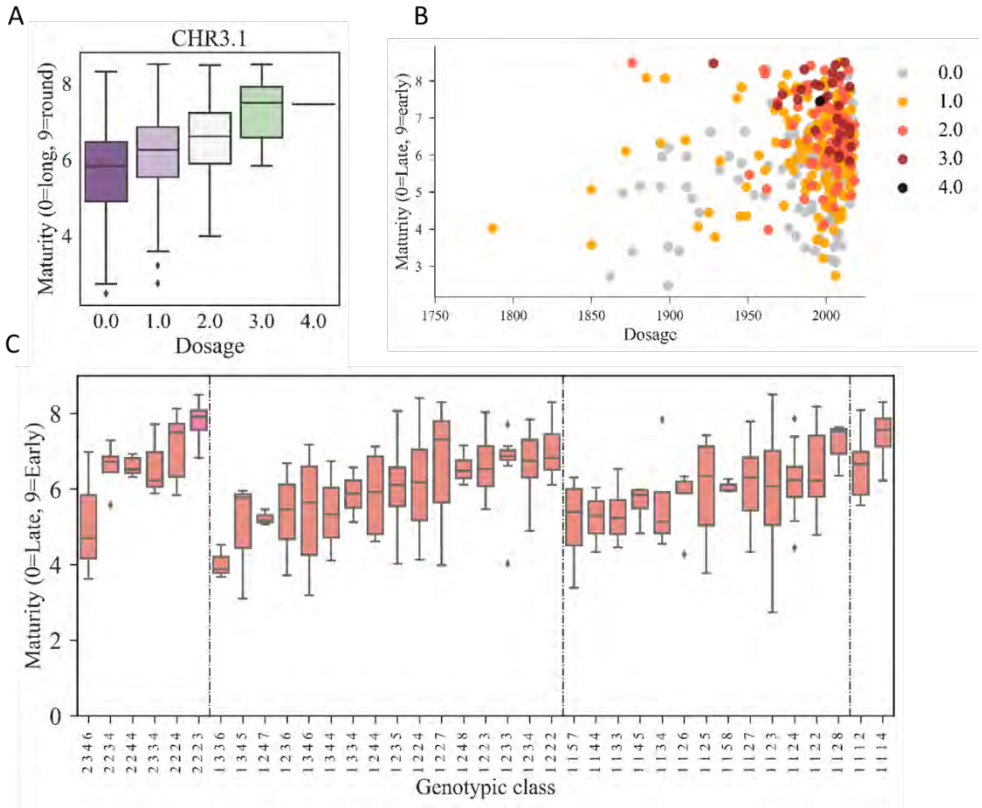


Figure 9. Haplotype analysis for chromosome 3 QTL for plant maturity (3 late and 9 early): A) Dosage effect of the significant haplotype on plant maturity, B) This haplotype is present in both old as well as newly introduced varieties. C) Average phenotype scores per class. The '1' represents the allele that is responsible for earliness. Absence of this allele leads to later tuberization.

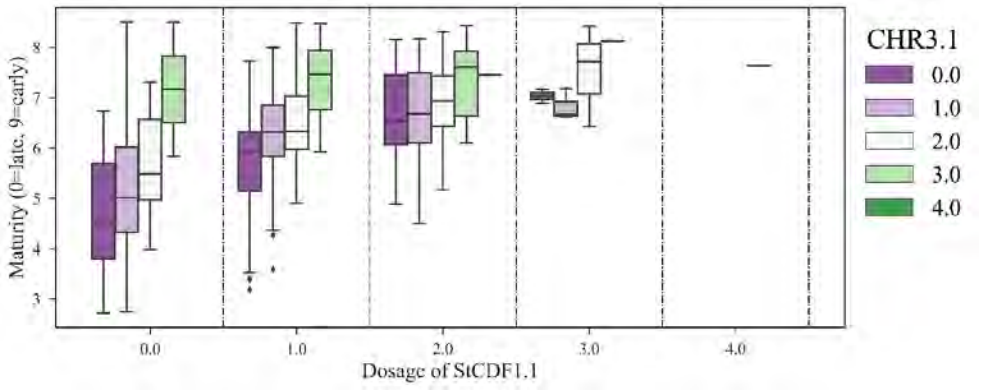


Figure 10. Distribution of trait values of Maturity for all genotypic classes observed for the joint presence of the CHR5 QTL *(StCDF1.A*) and chromosome 3 QTL.

Uniformity

We performed a GWAS for uniformity. As judged from Q-Q plots, the naive regression analysis showed high *p-value* inflation, and therefore only results for the kinship-corrected GWAS were reported. If a region is significant in either of the three different association mapping results, it is reported as a QTL. For each of these QTL regions we report the most significant SNP marker-trait association, and haplotype associated with the respective QTL (Table 4).

QTL	Model	SNP/block	chromosome	location	$-\log_{10}(p)$	Effect
I	single SNP	PotVar0041675	ST4.03ch01	74695176	4.10	-1.39
I	Multiple haplotype	block_1226	ST4.03ch01	76913651-76505319	3.92	
I	Allelic regression	block_1006	ST4.03ch01	70697501-70097516	4.08	-0.16
II	single SNP	PotVar0029766	ST4.03ch03	53739571	3.67	-0.31
II	Multiple haplotype	block_4551	ST4.03ch03	56604264-56501523	4.02	
II	Allelic regression	block_3632	ST4.03ch03	5430320-6354080	5.95	-0.54
III	single SNP	PotVar0111512	ST4.03ch04	67171388	2.91	-0.22
III	Multiple haplotype	block_5800	ST4.03ch04	66149755-66147380	4.33	
III	Allelic regression	block_5800	ST4.03ch04	66149755-66147380	5.03	-0.49
IV	single SNP	PotVar0073869	ST4.03ch06	57893874	4.51	-0.99
IV	Multiple haplotype	block_8370	ST4.03ch06	54084559-53989730	4.33	
IV	Allelic regression	block_8370	ST4.03ch06	54084559-53989730	5.21	-0.58
V	single SNP	PotVar0058084	ST4.03ch10		4.56	-0.63
V	Multiple haplotype	block_12114	ST4.03ch10	34303305-33324341	3.42	
V	Allelic regression	block_12210	ST4.03ch10	48863048-48717669	4.89	-0.35
VI	single SNP	PotVar0063968	ST4.03ch11	284050	2.89	0.17
VI	Multiple haplotype	block_12820	ST4.03ch11	7258676-7360405	4.03	
VI	Allelic regression	block_12799	ST4.03ch11	5532981-6513747	4.73	-0.28
VII	single SNP	PotVar0053743	ST4.03ch12	2200740	4.37	-0.30
VII	Multiple haplotype	block_13616	ST4.03ch12	1882025-1882952	3.04	
VII	Allelic regression	block_13627	ST4.03ch12	1951258-2200740	3.37	-0.34

Table 4. Overview of QTLs found for potato tuber uniformity. For each QTL the strongest association of each evaluated model is selected.

The Manhattan plot clearly identifies the complex genetic architecture of this trait (Figure 11). In total seven QTLs were defined based on a threshold of $-\log_{10}(p)$ of 4.0 (Table 4). Only the chromosome 6 QTL was identified in all three analysis. QTLs present

on chromosome 2, 3 and 11 were only found with haplotype-based analysis and not with single SNP association mapping. In contrast, the chromosome 12 QTL was only found using single SNP association mapping, and not with haplotype-association mapping.

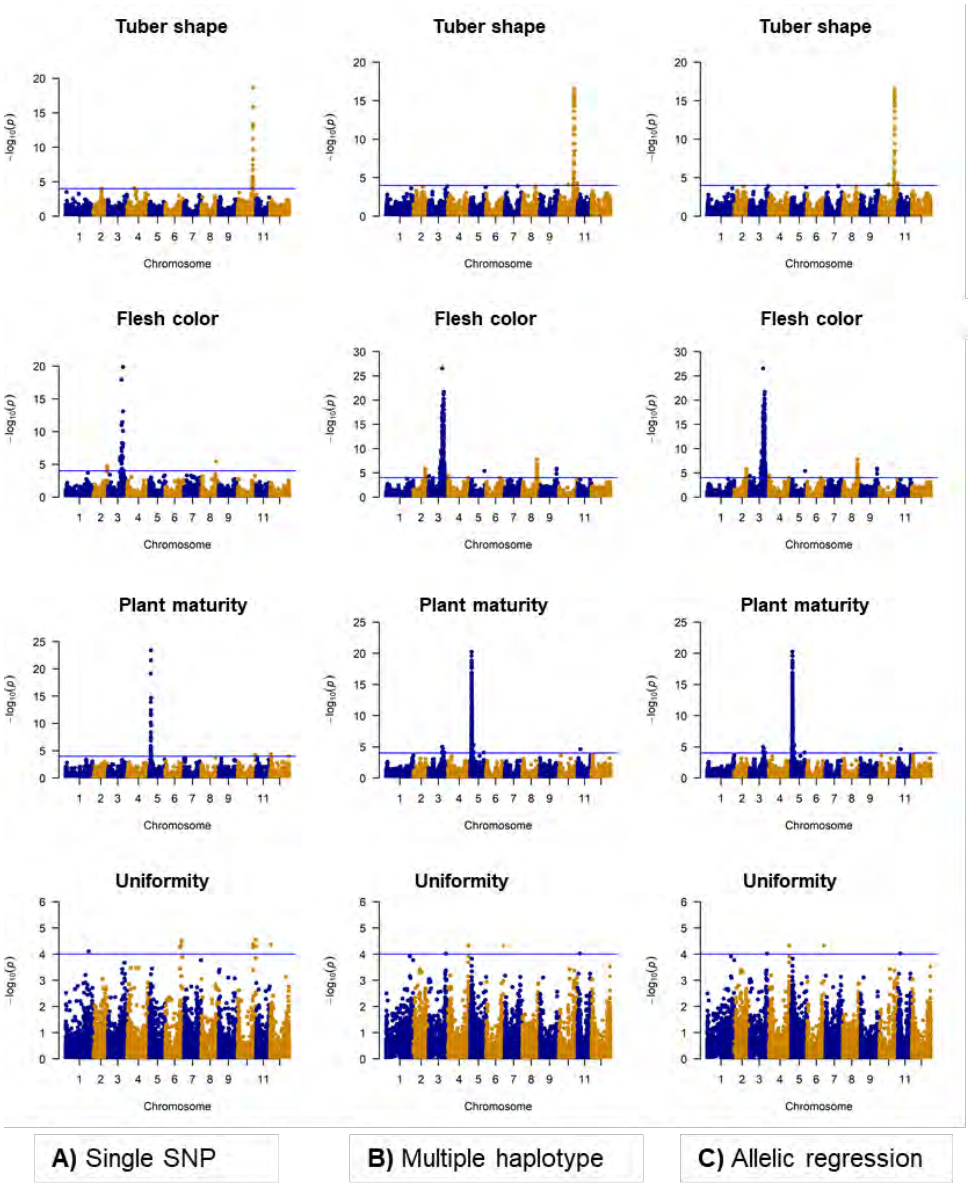


Figure 11. Manhattan plots of kinship-corrected association mapping, for tuber shape, flesh color, maturity and uniformity. The blue line is at the threshold of –log10(p) = 4. A) Single SNP analysis (N=14420). B) Multiple regression of all haplotypes (N=14409). C) Allelic regression of a single haplotype (N=122K).

Discussion

In this study, we report the comparison of haplotype-based genetic analysis to autotetraploid potato. Previously individual SNP markers were used to identify marker-trait associations for traits such as tuber shape, plant maturity (Kloosterman et al. 2013) and glycoalkaloid content (Vos et al. 2017). Here we were able to identify haplotypes with significant effect on phenotypic variation, whereas previous analyses have only identified bi-allelic SNP markers that are associated with phenotypic variation. A disadvantage of analyses with single SNPs, is the lack of knowledge of haplotype-specificity of each SNP-allele, which will be detrimental for the successful application of these markers in markers-assisted selection. From that perspective, the use of haplotypes as substitute for single SNP markers has clear advantages. So far a complicating factor was the reconstruction of haplotypes, which is challenging in polyploid crops such as potato. Previously we reported the development of a suitable approach for haplotype inference suitable for polyploid crops (Chapter 3). Application of this approach allowed us to reconstruct haplotypes of 10-SNP length, and subsequently use these haplotype blocks in association mapping of four traits within a potato variety panel, comprising 537 potato varieties.

In this study we investigated four traits with different underlying genetic architecture and performed both haplotype-based as single SNP association mapping. The found QTLs were to co-localize with results of previous studies, whenever such comparisons are possible. We first discuss our findings, and subsequently provide steps on how to use haplotype-based genetic analysis to improve the interpretation of association mapping experiments done in polyploid crops such as potato.

Association mapping confirms QTL positions

Flesh color

Arguably potato flesh color is a trait with a simple genetic architecture. Here we identified in agreement to earlier studies (Wolters et al. 2010; Bourke et al. 2018; Uitdewilligen et al. 2013) a single QTL, where presence of a single haplotype, allows to obtain varieties with yellow flesh color. This QTL is located in proximity of the *StCHY2* gene. In addition a minor effect QTL was found on chromosome 8, for which a putative candidate gene may be the *StCCD4* gene that was reported by Campbell et al. (2010).

Previously also the *StZEP1* gene was implicated in flesh color (Wolters et al. 2010), however here no significant association was found with SNPs or haplotypes in vicinity of the *ZEP* gene. In our variety panel we could not confirm the chromosome 9 QTL reported by Campbell (2014).

Plant maturity

For plant maturity we could confirm the major effect QTL reported by Kloosterman et al. (2013). Here we observe a single allele located in close proximity of this gene that explains nearly 24% of the phenotypic variation. In addition we identified a second QTL on chromosome 3 that explains 18% of the phenotypic variation. Between these two QTLs a significant interaction was found ($p < 0.01$), suggesting that differences in plant maturity are modulated by multiple loci. Surprisingly the chromosome 3 QTL was not observed with single SNP analysis and did not identify any association, but haplotype-based GWAS identified the significant association.

Tuber shape

For potato tuber shape, two QTLs were identified, from which one is the major QTL on chromosome 10 reported by (van Eck et al. 1994; Chapter 2) and the other QTL on chromosome 2 (Prashar et al. 2014). Previously we grouped highly correlated markers as identified with single SNP association mapping, to identify non-redundant associations (Willemsen et al. in preparation). As expected those correlations are coinciding with the haplotype structure at this locus. For instance PotVar0111687 has high correlation with solcap_snp_c2_25485, and lower correlation with solcap_snp_c1_8021. Indeed, both PotVar0111687 & solcap_snp_c2_25485 are present within all three round haplotypes (*Ro2-4*), whereas solcap_snp_c1_8021 is only present in one of these (*Ro2*). Here we would expect that PotVar0111687 has higher association than solcap_snp_c1_8021 as the former has higher predictability for the elongating allele (*Ro1*). Indeed, this was confirmed by our single SNP analysis, where the highest association is observed at PotVar0111687 $-\log_{10}(p)$ of 18.7), and solcap_snp_c1_8021 has a lower association signal of $-\log_{10}(p) = 5.86$.

Potato tuber uniformity

Arguably flesh color, tuber shape and plant maturity represent examples with simple genetic architecture, where major effect QTLs were easily identified. In contrast, potato

tuber uniformity represents a highly polygenic trait, for which minor-effect QTLs were observed (Figure 11, Table 3).

Evaluation of haplotype-based analysis

QTL discovery with haplotype-based association mapping

Numerous examples have demonstrated the power of GWAS for dissection of quantitative trait variation, both in diploids (Huang and Han, 2014), but also in polyploid crops such as potato (Rosyara, 2016; D'hoop 2014; Vos et al. 2017; Sharma et al. 2018). Nevertheless, it is known that bi-allelic SNPs present on SNP arrays inherently contain less information compared to previously used SSR marker systems, or amplicon-derived allele information (Schönhals et al. 2016; Schreiber et al. 2014). So far haplotype-based association mapping has been applied to many crops. For example, in maize multiple haplotypes were identified for a CO-like gene, responsible for differences in flowering (Yang et al. 2013). Also in other crops such as wheat (N' Daiye et al. 2017), soybean (Contreras-Soto et al. 2017) and oat (Bekele et al. 2018) successful application of genome-wide haplotype-based association analysis has led to the discovery of novel QTLs.

From a theoretical perspective haplotype-based association should result in higher power for QTL discovery (Schaid et al. 2002, Zaykin et al. 2002, Dudbridge, 2003), which in other words can be explained as the loss of power for QTL discovery due to unobserved haplotypes (Clark et al. 2004). Indeed several studies have shown that in general power for QTL detection increases if using haplotypes as substitute for single SNP markers (Akey et al. 2001, Morris and Kaplan, 2002), although these applications of haplotype-based GWAS are in human genetics, and might have little bearing on the results in outbred polyploid crops.

In our study the application of haplotype-based GWAS did overlap completely with single marker association mapping, with as exemption the detection of a novel QTL for plant maturity on chromosome 3, whereas for other traits (shape, flesh color) the same QTLs were detected. On basis of results for these traits, we can define several scenarios how application of haplotypes influence the results of association mapping.

Haplotype-specificity of SNP markers present at *StCHY2* for *StCHY2.1* allele

Firstly, in case of the chromosome 3 QTL for flesh color bi-allelic SNP markers were found that are specific for the *StCHY2.1* allele. In that case we expect a one-to-one relation between individual haplotype-specific markers and phenotypes, implying that usage of haplotypes does not improve the association analysis, which is confirmed by our analysis as all analysis (bi-allelic SNP, haplotype, allelic), resulted in similar strength of associations.

Lack of haplotype-specific SNPs for chromosome 3 QTL for maturity

Secondly, we identified a novel QTL for plant maturity on chromosome 3. With individual markers this QTL was not detected, but haplotype-based analysis identified a second QTL. Compared to the strong association of the chromosome 5 major effect QTL at the *StCDF1* gene, this QTL has a weaker association (R^2 = 18%). The most likely reason why this QTL escaped detection with single SNP analysis is the lack of haplotype-specific markers for this allele, suggesting that for this QTL haplotype-based regression has a clear benefit, due to higher distinguishing power of haplotypes compared to single SNP markers.

Tuber shape: Clear relation with haplotype specificity and significance.

Thirdly, for tuber shape a different scenario is observed, where the most significant SNP (PotVar0111687) groups all round alleles (*Ro2-4)*, allowing to disentangle the elongating allele (*Ro1*) from al round alleles. In contrast, other marker such as (solcap_snp_c1_8021), are specific to *Ro2*, and as a result have lower significance, clearly due to the lower haplotype-specificity of this bi-allelic SNP marker. Indeed, if breeders were to apply these markers for marker-assisted selection, knowledge about haplotype-specificity of individual SNP will allow to select markers that best tag existing alleles. One of the challenges that remain is the disentanglement of if all three alleles, conferring a round tuber shape, have different effects (Figure 7).

Polygenic traits are still difficult to interpret

Last but not least, most of the traits considered so far (plant maturity, flesh color & tuber shape) are traits with relatively simple genetic inheritance. As these examples might represent a oversimplified view of trait architecture, we also used potato tuber uniformity which displays a polygenic inheritance, with low broad-sense heritability (H^2 = 0.45). Both individual marker association mapping and haplotype-based association mapping

revealed several minor-effect QTLs (Figure 11, Table 3), which are only partly overlapping between the three analyses. As several QTLs were only detected with haplotype-based analysis and not in the individual marker association, and vice versa, the combination of all analysis allows to identify all six QTL regions, implying that the combined use of all three methods might be necessary to detect all QTLs.

Knowledge of haplotype composition is important for marker-assisted selection

In conclusion, the use of haplotype-based analysis has in general not led to the identification of multiple new QTLs. Nonetheless, the discovery of a novel QTL for plant maturity was only possible by using haplotype analysis. In addition we observe that knowledge of haplotype composition at a locus allows to improve the interpretation of genotypic data. In context of association this does not lead to a higher detection power, but from the perspective of a breeder this application of haplotypes should allow the selection of SNPs that are more reliable. An example of this is the major-effect QTL on chromosome 5 (located at the *StCDF1* gene) a single common allele was identified with strong positive effect on earliness (Figure 8). Most likely this allele represents the *StCDF1.2* or *StCDF1.3* allele as identified by Kloosterman et al. (2013). From the single SNP association mapping the most significantly associated SNP is PotVar79081, located 45kb upstream of the *StCDF1* gene. The haplotype data as generated here, allowed to determine that Potvar79081 is haplotype-specific for the for the earliness allele. In addition when examining the haplotype structure of markers present in close proximity –or within – the *StCDF1* gene no haplotype-specific markers can be found, suggesting that combinations of SNPs are needed to obtain the best correlation between trait and marker information.

Influence of genetic architecture

The lack of discovery of multiple novel QTLs might not be because we do not increase statistical power to detect QTLs, but might be coupled with the simple genetic architecture of these traits. Indeed, haplotype-based GWAS may result in higher power of QTL discovery in case of more complex traits, as suggested by Hamblin et al. (2011). Here, in case of potato tuber uniformity, we observed that haplotype-based GWAS discovered most QTLs in both analyses, whereas several QTLs were detected in either single SNP analysis or haplotype-based analysis. Based on the lack of coherence signals

between discovered QTLs between these analysis would suggest that both analysis need to be combined, to give a comprehensive overview of all putative QTL locations.

Specifics of the SOL-STW panel

Here we used the SOL-STW panel which deliberately was designed to capture the wide genetic variation of the commercial potato genepool (Uitdewilligen et al. 2013, Vos et al. 2015), and therefore has a good representation of markers with low allele frequency (more likely to be haplotype-specific) and an under representation of markers with relatively high frequency (more likely to be present in alleles that are common/ not be haplotype-specific), suggesting that a large proportion of the SNP markers that were used here are largely haplotype-specific. The application of haplotype-based regression analysis might yield more results for SNP arrays that lack such haplotype-specific SNPs. For instance, the 8303 Illumina Infinium SNP array (Felcher et al. 2012), displays a different allele frequency distribution, where the majority of SNPs have minor allele frequencies higher than 10% (Sharma et al. 2018), suggesting that less haplotype-specific markers are present on this array.

Genotyping-by-sequencing

In future, whole-genome sequencing data will rapidly become available, which might contain all causative variants that influence a trait, which likely will be specific for the causative allele. However, even if all genetic variation present in a population is known, the capricious relation between LD and neutral versus causative variants still remains (Korte and Farlow 2013). In the end, the frequency of each individual haplotype is determined by the demographic history of the population, where for instance if a haplotype represents old 'standing' variation, it is more likely that recombination has produced recombinant alleles with moderate frequencies. If no or limited, markers are present within a small distance of the causative variant, recombinant alleles are expected to contain less haplotype-specific SNPs, with as consequence a loss of power to detect any significant allele-phenotype association. The opposite might be true for introgression segments, where often SNPs are introgressed that are previously not found in the genepool, allowing to rapidly find association with a phenotype.

Reconstruction accuracy influences the results of haplotype-based analysis

One aspect that plays a role in haplotype-based analysis is the reconstruction certainty (Stram and Seshan, 2012). Most likely genotyping error can generate false associations, especially for SNP markers with low minor allele frequency. Previously it was found that marker scores from the SOL-STW array, as used here, are highly accurate, with 0.02% errors were found between replicated diploid samples (Vos et al. 2016). In contrast existing methodologies for haplotype assignment often report an error percentage between 5-10% (Neigenfind et al. 2013; Shen et al. 2016; Chapter 4). Previously we determined that for the approach introduced in Chapter 4, most phasings errors can be discarded using a simple cut-off in allele frequency, as reconstruction error mainly results in low frequent haplotypes. These errors will mainly in the varieties, where haplotypes are considered absent, but in reality they are present (i.e. there is loss of heterozygosity in the direction of the major allele). The necessity of accurate phase reconstruction is also shown in QTL mapping studies in bi-parental populations, as the result of IBD-based QTL mapping is greatly influenced by the genotypic information content (Bourke et al. 2018).

Prospects and future application of haplotype-based analysis.

The application of association mapping to crops has shed light on the genetic architecture of many important agronomical traits. Here, we explore a new layer of information, which allows to link trait variation directly with allelic diversity present at the QTL. Addition of haplotype-based genetic analysis to the repertoire of the geneticist will allow to increase the QTL detection power, by increasing the one-to-one relation that is expected between allele and phenotype. For the routine application of haplotype-based association mapping the availability of fast and accurate phasing algorithms is necessary. Given that the application of genotyping-by-sequencing data is rapidly increasing, further research is needed into optimizing existing phase inference methods to cope with these data sources. Likewise more insight is needed into adequate statistical methods for association mapping within polyploid species. As haplotype inference will lead to an addition of uncertainty to genotypic data, we propose to use haplotype-based association analysis in combination with single marker analysis. The knowledge of haplotypes will improve the interpretation of QTLs, and allows potato breeders to

optimize selection of haplotype-specific markers, and apply these in marker-assisted selection.

Acknowledgments

JHW is supported by a grant of the Dutch Science Organisation NWO (project 831.14.002) The potato breeding companies Averis Seeds B.V., HZPC Holland B.V., KWS POTATO B.V. and Meijer B.V. are acknowledged for contributing phenotypic data. The Dutch Technology Foundation (STW grant WPB-7926) financed the development of the SOL-STW array and SNP-data production.

Additional files

File S1: Q-Q plots of GWAS.

File S2: Results of Multi-locus regression flesh colour.

File S3: Results of Multi-locus regression maturity.

Chapter 6

Poly-Imputer assigns haplotypes to unphased genotype data

Johan H. Willemsen, Richard G. F. Visser & Herman J. van Eck

Abstract

Poly-Imputer is a tool to assign known haplotypes to individuals for which only unphased genotypic data is available. In polyploids, SNP phasing is challenging and requires extensive computational resources. However, if a reference library of high-quality curated haplotypes are known, disentangling of genotypic data of one individual into their respective haplotypes becomes trivial. While assigning these haplotypes to individual samples, missing data and errors in allele dosages can be corrected, resulting in more complete and accurate haplotypes. We developed a method to assign reference haplotypes to un-phased genotypic data. Here we described three applications of this tool: 1) Impute parental haplotypes to a segregating full-sub population, 2) Assign haplotypes in a panel of unrelated potato varieties, based on sequencing data haplotypes. 3) Refine and curate haplotypes that are obtained by haplotype inference. The application of this tool allows to quickly screen additional samples for the occurrence of haplotypes that are present in a reference haplotype library.

Keywords

Solanum tuberosum, haplotype imputation, polyploids, phasing

Introduction

For outbreeding autopolyploid crops such as potato (Vos et al. 2015; Hamilton, 2011; Felcher et al. 2012) chrysanthemum (van Geest et al. 2017) and rose (Vukosavljev et al. 2016) high-density SNP genotyping platforms are being adopted quickly. This is coupled with the development of tools to accommodate these high marker densities for use in linkage map construction (Hackett et al. 2003; Bourke et al. 2016), QTL mapping (Bourke et al. 2018), or genome-wide association studies (Rosyara et al. 2016). Likewise, in recent years a shift can be seen from the use of bi-allelic markers towards the adoption of haplotype markers in genetic analysis. A single bi-allelic marker offers incomplete information about the allele composition in case of multiple alleles in a polyploid individual. The use of multiple bi-allelic SNP markers will allow to reconstruct haplotypes and enable researchers to obtain full classification of a locus.

Much progress has been made in recent years in reconstructing haplotypes from either sequencing data (Aguiar et al. 2013; Berger et al. 2014; Chapter 3), or by using statistical phasing methodology (Neigenfind et al. 2008; Su et al. 2008; Shen et al. 2016, Chapter 4). A downside of the application of these methods is the large computational requirements, especially if a genotype panel comprises many individuals, or large numbers of markers. Also, every time new samples are genotyped, phasing needs to be repeated, whereas in most cases no new haplotypes will be discovered. In case of 'unrelated' material, genotyping errors are difficult to spot, further complicating phasing efforts. In some cases, pre-existing haplotype information is already available. For instance parental phasing in a full-sib population is achieved by linkage mapping, providing knowledge over segregating alleles, or when inbred individuals are genotyped.

A practical solution to obtain accurate inference of haplotypes, without much effort, would be to exploit the use of a library of curated high-quality haplotypes and to establish the allelic configuration of additional unphased samples. The simplest approach to achieve this is with use of haplotype-specific or tag-SNPs. These bi-allelic SNPs are known to distinguish a single haplotype from all other haplotypes. This allows identification of haplotypes on basis of a single SNP, that acts as proxy for the complete haplotype (Johnson et al. 2001). Nevertheless, in many cases a single SNP will not uniquely tag only one allele, is not specific for other haplotypes. To recognise other haplotypes, the use of multiple (partially) haplotype-specific SNPs is required.

Many approaches have been developed that perform haplotype phasing and imputation in diploid organisms. In spite of these developments, most approaches cannot be applied directly to autopolyploid crops such as potato. Generally, these imputation methods are divided into two categories: tools that make use of linkage disequilibrium information and those using pedigree and or linkage information. The first category includes tools such as fastPHASE (Sheet and Stephens, 2006), MacH (Li et al. 2010), Beagle (Browning and Browning 2009) and Impute2 (Howie et al. 2009), whereas the second category includes plantImpute (Hickey et al. 2015) and pediHaplotyper (Voorrips et al. 2016).

To our knowledge no publicly available algorithm has been explicitly designed for haplotype imputation in polyploid crops, although studies have used haplotype imputation by use of individual tag-SNPS (Uitdewilligen et al. 2012, Wolters et al. 2010) to disentangle the allelic configuration of tetraploid potato cultivars. The use of these imputation strategies would potentially allow minimization of missing genotype data and therefore improve QTL detection accuracy, as demonstrated by results observed in pigs (Hickey et al. 2011, Hickey et al. 2012), maize (Huang et al. 2014) and tomato (van Binsbergen et al. 2014).

To identify haplotypes, conventionally haplotype assembly (e.g. haplotype reconstruction based on sequencing data), or haplotype inference is applied (e.g. haplotype reconstruction based on SNP genotyping data). The output haplotypes will contain a mixture of correctly reconstructed haplotypes and erroneous haplotypes, which need to be filtered and curated. In case of short read sequencing-based haplotype assembly, the result is often a fragmented haplotype assembly, where full haplotype information is only available for a small proportion of all haplotyped varieties (Chapter 3). Likewise, haplotype inference depends crucially on the genotype accuracy of input data, and the composition of the genotyped samples. Moreover, in case of dosage errors, erroneous haplotypes will be reconstructed (Chapter 4). Subsequent downstream genetic analysis needs rigorous post-processing of haplotypes to avoid the use of erroneous haplotypes.

Approach

The need for complete, high-quality haplotype information in genetic studies, irrespective of whether haplotypes originated from sequencing-based haplotype reconstruction, or application of haplotype inference to genotypic data of seemingly

unrelated varieties, has motivated us to develop a tool that allows to impute haplotypes based on a set of high-quality reference haplotypes. In general the procedure involves three steps: 1) For a set of SNPs haplotype reconstruction is performed, 2) Manual curation of these haplotypes based on pedigree relations, allele frequency and additional quality criteria and reconstruction of a reference haplotype library, 3) Haplotype imputation in other individuals that for which SNP phasing failed or resulted in partial phasings. In this paper, we do not provide tools to perform step 1 and 2, but describe a method to achieve haplotype imputation (step 3).

Material and methods

As input for our algorithm a pre-existing haplotype library and genotypic data of each individual are used. The complete method employs three steps: (1) Determination of possible haplotype configurations given the set of reference haplotypes. (2) Determination of allelic configuration that are consistent with genotypic data of a single individual (3) Assignment of specific haplotypes to each individual.

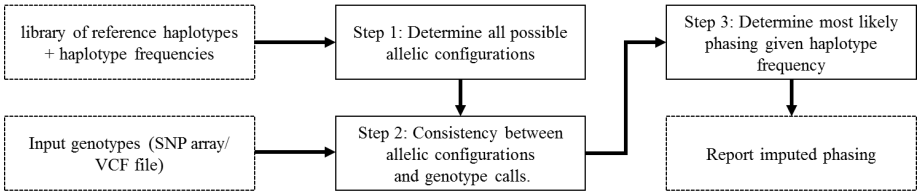


Figure 1. Schematic overview of Poly-Imputer approach.

Step 1: Determining possible allelic configurations.

Before imputation we determine all possible allelic configurations that can occur in each individual given the set of input haplotypes (i.e. all combinations of four haplotypes out of all haplotypes in case of tetraploids) and ploidy level. For each of these allelic configurations we calculate the expected genotype frequency using the assumption of random mating. For n distinct alleles in tetraploids, the expected genotype frequencies for each allelic configuration can be calculated under assumption of Hardy-Weinberg equilibrium by the individual terms in the multinomial expansion of $(p_1 + \cdots + p_n)^k$. In case the assumption of random mating is not valid, for the selected population, we assume that each phasing configuration has equal expected genotype frequency and gets assigned equal probability.

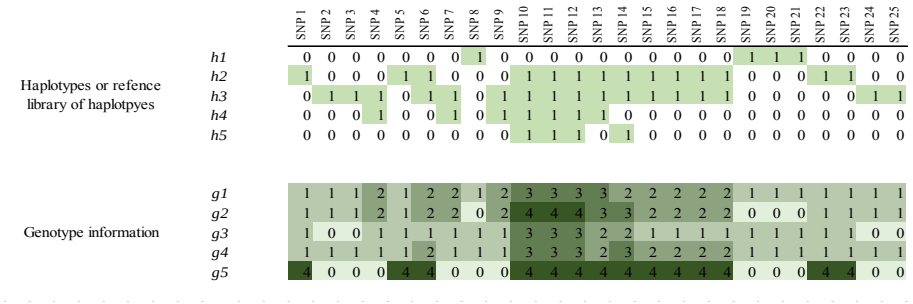


Figure 2. Typical imputation scenario where a library of reference haplotypes (*h1-h5)* is projected on genotyping data (*g1-g5*) The joint presence of multiple SNPs allows to identify which combination of haplotypes is present in each sample.

Step 2: Calculating of consistency between a allelic configuration and observed allele dosages.

After determination of the set of possible haplotype configurations and their expected frequency in the population, we score for each individual which haplotype configurations can explain the dosages of the individual SNPs. For instance, a two-locus genotype containing two simplex SNPs at each locus (11) can be disentangled into a phasing configuration 11|00|00|00 or 10|01|00|00, as summation of the allele dosages of individual SNPs for both allelic configurations results in 11. Therefore, a simple approach to determine consistency is to compute a genotype vector for each allelic configuration (*g'*), and compare this with the observed allele dosages of a single individual (*g)*. The comparison of these scores allows to score consistency (e.g. number of mismatches) between an allelic configuration and a genotype by computing the hamming distance. The simultaneous presence of multiple (partial) haplotype-specific SNPs allows to determine the best allelic configuration explaining the pattern of dosages observed in a single genotype.

An example of calculating this consistency measure is given in Figure 2. A segment of 25 adjacent genotyped SNPs (*g1)* is compared to the summation of dosages of a theoretical phasing configuration, consisting out of *h1-h4*. Summation of dosage of this haplotype configuration results in a genotype (*g').* In this case, in the whole segment of 25 SNPs a hamming distance of 0.95 (21/22) is observed between *g'* and *g*, as only SNP4 has a mismatch. Note that missing values might be present at any of the genotyped SNPs (*NA).* A missing value present in a segment of SNPs is not used for calculating consistency.

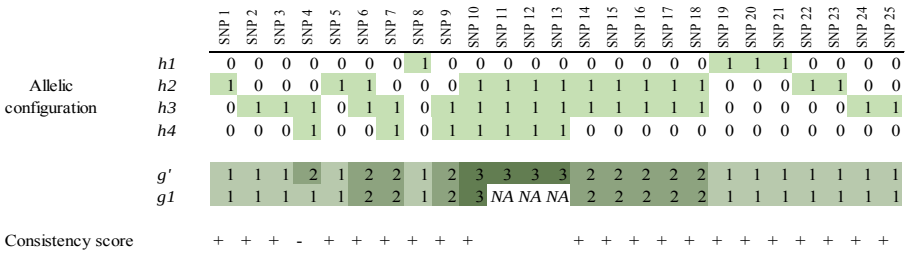


Figure 3. Determination of consistency between marker dosages and an allelic configuration of multiple reference haplotypes. A single haplotype configuration (*h1-h4)* is checked for consistency with dosage data obtained in an individual (*g1).* The + represents a consistent dosage score between *g'* and *g1.* The – represents an inconsistent dosage score. Here consistency is ranked 0.95 (21/22). In case there are missing values (*NA*) we discard these SNPs.

One of the advantages of haplotype imputation is the ability to also impute missing genotype calls, provided that most SNPs in an interval are genotyped. The downside of a high amount of missing genotype calls is that imputation will become less reliable, which can be mediated by specifying the maximum amount of missing data that is allowed during imputation. Here we determine the consistency score for each allelic configuration we calculate the consistency score. If there are no consistent solutions, imputation does not proceed.

Step 3: Assigning the most likely solution to un-phased individuals.

At this point, we have for each individual a set of likely allelic configurations, but still need to assign the best solution to each individual. In some cases a single configuration might be consistent with the genotype data (e.g. genotype 410 can only originate from configuration 110|100|100|100), which can be reported directly. In other cases more than one allelic configuration are consistent with dosage data, where multiple combinations of haplotype result in the same genotype vector. The determination which allelic configuration is the most likely is subsequently decided by using the expected genotype frequencies, which are calculated by the underlying haplotype frequencies. If no population-allele frequencies are known, we either report the solution with the highest consistency, or produce output with all consistent phasings.

As our method for imputation does not generate new haplotypes, but rather assigns known haplotypes to unphased genotypes, a haplotype assignment might conflict with genotypic data. This can occur because of several reasons: (1) Recombination can lead to alleles not present in the current set of reference haplotypes, (2) Dosage errors will lead

to erroneous genotype vectors, and therefore result in faulty haplotype assignment. For instance, if genotype vector 11 contains a dosage error at the first SNP, resulting in genotype 21, the allelic configurations that are in agreement with are 10|10|01|00 and 11|10|00|00 instead of 11|00|00|00 or 10|01|00|00.

Implementation

Poly-Imputer is implemented in Python2.7 and can be executed on any system for which Python is available. Haplotype imputation is handled by a command-line script (File S1). The formats of the input and output files are detailed in the readme file and example dataset.

Datasets

F1 population: The offspring of a full-sib population with 233 individuals, previously genotyped with a SNP array and originating from the cross between 'Altus' and 'Colomba' was used for imputation. This population was previously used to generate a high-density linkage map (Bourke et al. 2016). For each progeny linkage phase was reconstructed using tetraorigin (Zheng et al. 2016). Here we binned all segregating markers in windows of 1 cM, and in each interval haplotype imputation was performed. Expected haplotype frequencies were obtained by using the counts of unique haplotype across the two parents (i.e. SxS = 0.25, SxN = 0.125, and so forth).

Haplotypes of the StGWD1 locus: A library of reference haplotypes were previously reconstructed using the sequencing reads of Uitdewilligen et al. (2013) for a region of 300 bp comprising 18 SNPs (Chapter 3). Imputation was performed using sequencing-based dosages obtained with Freebayes (Garrison et al. 2008).

Haplotypes of a potato variety panel: Previously haplotypes were generated using the haplotype inference method described in chapter 4, over a sliding window of 10 SNPs, resulting in 14K intervals. Each of the intervals was used for association mapping (Chapter 5). For the purpose of this study we extracted the haplotype block with the most significant association to plant maturity (interval 6650). From this interval with 10 SNPs, 31 unique haplotypes were extracted and allele frequencies were calculated.

Results

Application 1: Identifying alleles in progeny of a bi-parental F1 population

In this study, we used genotype data collected with a SNP array, from a tetraploid bi-parental F1 mapping population to assess the allele assignment using Poly-Imputer. In a previous study, all eight homologous chromosomes were reconstructed with use of conventional linkage mapping. In the progeny phasing was achieved with tetraOrigin (Zheng et al. 2016). To allow the comparison of imputed haplotypes with tetraOrigin based IBD probabilities (Bourke et al. 2018), these probabilities were converted to discrete haplotype information. As Poly-Imputer only works in absence of recombination we divided all twelve potato chromosomes in bins of 1 cM. In each of these 1cM intervals (n=644), we employed Poly-Imputer and assigned which parental haplotypes were transmitted to each of the 235 progeny. Comparison with phasing results from tetraOrigin showed an high concordance of on average 98% between phasings, which increases with considering more SNPs in an interval (See Figure 4A).

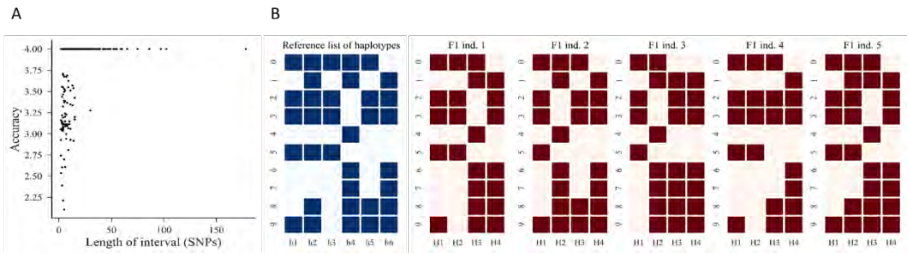


Figure 4. A) Imputation accuracy in 664 intervals of 1 cM over 12 potato chromosomes in the bi-parental AxC population. B) Example of a haplotype comprising a library of 6 haplotypes over 10 SNPs.

Application 2: Identifying alleles from GBS dosage calls.

Previously we assembled haplotypes using sequencing reads for 2400 regions in 800 potato genes in 83 tetraploid potato varieties with short-read Illumina sequencing data (*Chapter 3*). This resulted in a fragmented assembly, as a single region is often broken into multiple discontinuous haplotype blocks. For instance, we determined the haplotype structure for a 300bp fragment of the *StGWD1* gene, where over all informative varieties, 12 unique haplotypes were reconstructed (Figure 5A). However, from the total of 83 varieties, 14% did not contain full-length haplotypes and were split into multiple disjoint haplotype blocks. Between these adjacent haplotype blocks, no linkage phase is known. Here we use the 12 complete haplotypes that were reconstructed

in 63 varieties as reference haplotypes and apply Poly-Imputer to identify which haplotypes are present in varieties that contain partial haplotype information. To judge the quality of imputation, the concordance between full-length haplotype solutions in the remainder of varieties, and imputed solutions was calculated, resulting in a concordance of 98%. The 2% erroneous imputation can occur due to errors in dosage of individual SNPs. After imputation, the allele composition of 80 varieties could be determined (File S3). Three varieties were not imputed, as multiple allelic configuration had equal support.

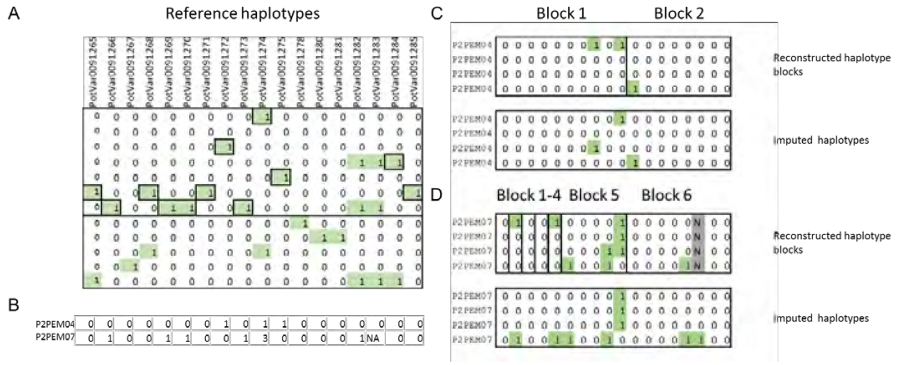


Figure 5. Example of haplotype imputation for haplotype blocks present in the *StGWD1* gene. A) List of haplotypes obtained from 71 varieties. B) Genotypes of variety Kerpondy (P2PEM04) and entland Dell (P2PEM07). C) Imputation in variety Kerpondy D) Imputation in variety Pentland Dell.

Application 3: Refine haplotypes

In a previous study haplotypes were reconstructed over a sliding window of 10 SNPs, using genotypic data of 14K usable SNP markers from 537 varieties (Chapter 5).These haplotypes still contain missing data and erroneous haplotypes. Phasing errors often occur by loss of heterozygosity in the direction of the major allele (e.g. an allele is not present where it should be present). To demonstrate the improvement that can be achieved by application of haplotype imputation we refined the haplotype composition at a single block of 10 SNP markers (PGSC4.03 CHR05: 4488075-4489590). Within this position a total of 31 unique haplotypes were found, with frequencies ranging from 0.005% to 27.5%, but only one haplotype showed a significant association to potato plant maturity. In addition 16 out of 537 genotypes were not haplotyped as individual SNP dosages were missing.

We applied Poly-Imputer by using the library of 31 reference haplotypes as input and imputed these haplotypes in each of the 537 varieties (File S3). Previously we determined that the causative allele had a allele frequency of 0.27, which after imputation increased to 0.29. In addition we observe that 13 out of 31 haplotypes were removed after imputation. These unused haplotypes had frequencies ranging from 0.0005-0.01 before imputation and likely represent erroneous haplotypes. In terms of allele composition, approximately 429 genotypes had the same allele composition as before imputation, whereas 106 genotypes were different after imputation.

The use of imputed haplotypes in an association analysis allowed to revisit the association of haplotypes with plant maturity. Here, the association with plant maturity slightly increased from a $-\log_{10}(P)$ of 38 to a $-\log_{10}(P)$ of 40.6 using the imputed haplotypes. The explained phenotypic variation by this association increased from 27.1% to 28.9%, showing only a marginal improvement. Nevertheless the use of haplotype imputation to be preferred as it imputes missing data. Here, for the 14 out of the 16 genotypes with missing genotypic values at one or more SNPS, we could determine the configuration of haplotypes. For instance, previously we could not determine the occurrence of the early haplotype in a progenitor clone (P8WUR045), where a missing value was observed at PotVar0079086. Clearly, this variety contains PotVar0079038 and PotVar0079081, indicative of the presence of the early allele and this variety has an early maturity type. This suggested that this variety should contain the early allele. Once imputation was performed, the imputation allowed to resolve the missing dosage score at PotVar0079081.

Discussion

Haplotype reconstruction errors, whether or not caused by dosage errors or the quality of sequencing data, seriously hamper the application of haplotype data in genetic studies. For successful application of haplotypes in genetic studies, these errors have to be removed. We developed an approach that uses a curated set of reference haplotypes and performs haplotype imputation in genotyped material, which has not undergone haplotype detection, or in which only partial genotype information is present. Haplotype reconstruction requires considerable computational resources, but application of Poly-Imputer allows to quickly screen large numbers of samples and determine their

haplotype composition. Our approach is based on predicting which haplotypes are present, based on dosage information of (multiple) individual SNP markers.

The haplotype imputation approach as used in this study requires the availability of a library of reference haplotypes, which needs to be produced with other software. In this study we made use of existing haplotype data that was generated using sequencing data (Chapter 3), or by means of haplotype inference on bi-allelic SNP markers obtained from SNP array data (Chapter 4, 5). In contrast, phasing in the tetraploid full-sib population was achieved by traditional linkage mapping (Bourke et al. 2016; Bourke et al. 2018), followed by phase reconstruction within each tetraploid progeny (Zheng et al. 2016). The accuracy of haplotype imputation depends on how well these original haplotypes were estimated, which might depend on marker density, relatedness, sample size and demographic history of each allele. Any imperfection in the haplotype library limits the accuracy of imputation, suggesting that the reference haplotype library needs to be carefully reconstructed. A drawback of our proposed method is the assumption that there is no recombination within a single haploblock. If a recombinant allele is present, an erroneous assignment will be performed, or no assignment at all.

Arguably our approach does not allow to generate new haplotypes, and depends on the assumption that an initial reference panel contained most or all occurring allelic variation. Increasing the panel size will soon result in a diminishing return as most of the genetic diversity is already observed in a limited set of cultivars. This is exemplified by results of Uitdewilligen et al. (2013), where it was estimated that after sequencing a random subset of 20 cultivars most genetic variation was already discovered (See Figure 2 Uitdewilligen et al. 2013). In fact even in small panels a haplotype with allele frequency of 1% can be found (see Figure 6).

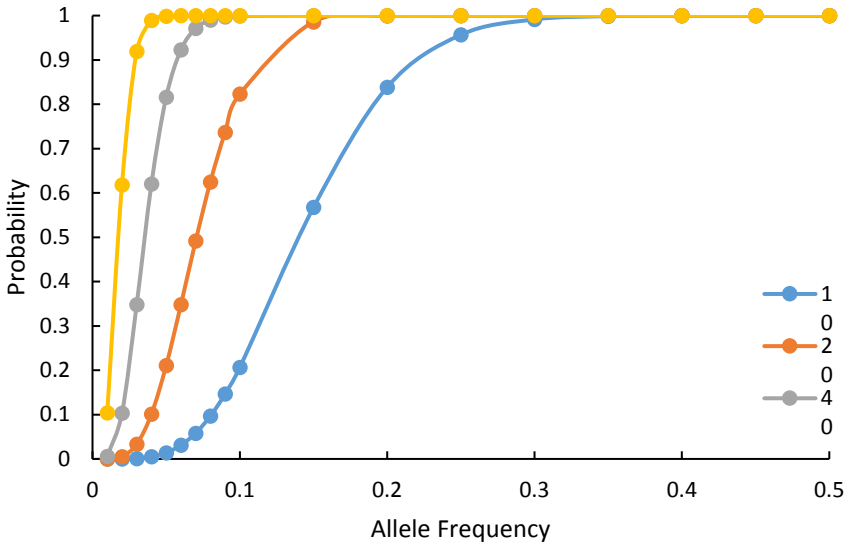


Figure 6. Probability of observing a haplotype with varying allele frequency at least one time in panels of 10, 20, 40 and 80 tetraploid varieties. Probabilities are calculated using a binomial model.

In this study, we applied Poly-Imputer to 231 full-sib offspring of a cross between two tetraploid potato varieties. As parental haplotypes are phased during linkage mapping (Bourke et al. 2016), and for each segment of 1 cM, the transmission of alleles to offspring is known, the expected allele frequencies across the 231 progeny could be calculated. We compared these results to the ground-truth as provided by tetraOrigin-based phases in all progenies. In all 644 windows of 1 cM, this resulted in an agreement of 98%, suggesting that our approach is able to reliably assign these haplotypes to each descendent. In addition, we also employed haplotype imputation to a haploblock comprising 10 SNPs located in close proximity to the *StCDF1* gene, and revisited existing haplotype-trait association to plant maturity. The application of Poly-Imputer allowed to refine existing haplotype solutions and impute missing genotype calls. With respect to haplotype phasing, most errors can be detected by using a allele frequency cut-off (i.e. errors are not systematic), or by selecting varieties for which haplotypes are likely to be true (e.g In agreement with IBD). Application of Poly-Imputer could lead to stronger QTL detection power if performed over all reconstructed haplotypes.

Improving sequencing-based haplotypes

One of the main reasons for the development of this haplotype imputation approach is the discontinuity in haplotype blocks obtained from sequencing-based haplotype reconstruction. Any single individual haplotype assembly method will result in full-length haplotypes for only in a subset of all varieties. Each of the those varieties, could contain good quality variants, while haplotype reconstruction results in multiple discontinuous haplotype blocks. Between these blocks linkage phase is unknown (Figure 4B). To test whether haplotype imputation can improve the contiguity of sequencing-based haplotypes, we reconstructed haplotypes in a segment of 300 bp of the *StGWD1* gene. Before imputation for 12% of the varieties no haplotypes were known. However, after imputation almost all varieties were assigned four haplotypes. The high concordance between the non-imputed haplotypes and imputed haplotypes (98%), suggests that Poly-Imputer is successful in improving the contiguity of the haplotype assembly and can be used in conjunction with haplotype assembly methods.

What is next for Poly-Imputer?

At its core Poly-Imputer executes a simple routine which involves determining all possible haplotype configurations (Step 1), followed by the comparison of the sum of dosages of this configuration, to the observed dosage calls (Step 3). Subsequently, the best haplotype configuration is assigned to each sample, given the underlying allele frequencies. One of the advantages of Poly-Imputer is that it is not dependent on population structure or availability of pedigree data. If such data is known, these genetic relations may constrain the solutions for haplotype imputation. Arguably this study present only anecdotal examples of possible applications of haplotype imputation, and a more rigorous evaluation of factors influencing imputation accuracy are needed.

Conclusion

In conclusion, Poly-Imputer can be applied to identify haplotypes using unphased genotypic data from both sequencing-based genotyping or dosage calls originating from SNP arrays (Application 1,3). In addition, it also can be used to improve contiguity of sequencing-based haplotype blocks (Application 2). In combination with any phasing software this tool will allow refining existing haplotypes, and quickly screen un-phased individuals for the presence of these haplotypes. The routine application of Poly-Imputer

could decrease the computational resources that are needed for building a complete haplotype map of any polyploid crop.

Acknowledgements

The authors wish to thank Dr. P.M. Bourke and Dr. Ir. C.A. Maliepaard for providing the tetra-Origin based phasing information of the tetraploid F1 population (A×C). JHW is supported by a grant of the Dutch Science Organisation NWO (project 831.14.002). The Dutch Technology Foundation (STW grant WPB-7926) financed the development of the SolSTW array and SNP-data production.

Additional files

File S1: Package to perform imputation.

File S2: Haplotype composition of sequencing-based *StGWD1* haplotypes.

File S3: Haplotype composition of 10-SNP length haplotype block at *StCDF1* gene.

Chapter 7

Haplotype diversity at the *StCDF1* gene and quantification of the effect on maturity in potato

Johan H. Willemsen, Jan Uitdewilligen, Richard G. F. Visser, Herman J. van Eck.

Abstract

Large differences in potato plant maturity are mediated by a major effect QTL on chromosome *5*. Previously the causative gene *StCDF1* was mapped and three alleles were found in a diploid F1 population: *StCDF1.1, StCDF1.2, StCDF1.3.* The *StCDF1.3* allele contained a 865bp transposon insertion, whereas the *StCDF1.2* allele contained a 7bp-footprint originating from a transposon excision event. Both mutations cause a truncated protein due to a premature stop codon and act as a dominant early allele unable to delay tuberization until short day lengths. In this study we determined the intra-gene allele diversity of a part of the *StCDF1* gene, using haplotypes originating from next-generation sequencing data from 83 potato varieties. We observed that one allele has a significant effect on early plant maturity. The presence of the transposon and/or footprint allele was recorded for each variety. Association analysis between these haplotypes and plant maturity verified the significance of *StCDF1.3* transposon allele on early maturity, but failed to identify a significant effect of the footprint allele (*StCDF1.2).*

Keywords:

Solanum tuberosum, haplotype, allelic diversity, maturity, tuberization.

Introduction

Potato plant maturity is considered a quantitative trait for which a major effect QTL located on potato chromosome *5* modulates differences in plant maturity, and thus day-light dependent tuberization. Early maturity allows the cultivation of the short-day potato under long-day conditions. Different market segments require different maturity classes, which phenotype is generally recorded by breeders by observing foliage senescence, termination of flowering, prostrate stems, and/or inactive apical meristems as a proxy for tuberization. Previously this major effect QTL was mapped to the *StCDF1* gene, PGSC0003DMG400018407 (Kloosterman et al. 2013). A diploid mapping population segregating for plant maturity (Visker et al, 2003) allowed to associate late maturity to homozygosity for the wild type *StCDF1.1* allele. This functional allele shows 100% sequence identity with the *StCDF1* gene of the DM reference genome. Early maturity was associated with the presence of the allelic variants *StCDF1.2* and/or *StCDF1.3*.

Genomic information present in sequenced BAC clones of RH-89-039-16 allowed to determine that the *StCDF1.3* allele contained a transposon insertion, whereas the *StCDF1.2* allele displays a transposon excision event, where only a footprint (TSD; Target Site Duplication) of 7bp remained. In the *StCDF1.2* allele, this insertion results in a frame-shift introducing a premature stop codon. Both these alleles lead to an early phenotype. The molecular characterization of these alleles showed that the lack of the terminal domain III allows these proteins to evade post-translational degradation by the complex of *FKF1-GI* light receptors. An accumulation of the *StCDF1* protein subsequently leads to an early phenotype.

So far the allelic variation present at this locus has not been characterized in commercial potato germplasm. Previous reports have only characterized allelic variation in a number of primitive landraces, which comprises only few tetraploid potato genotypes (Hardigan et al. 2017), but failed to assess allele diversity within a large set of tetraploid potato genotypes. To facilitate marker-assisted breeding we describe a catalogue of different *StCDF1* haplotypes, their allele frequency, and extent to which this allelic variation explains the differences in maturity. The results in this study show that the *StCDF1* displays a large number of haplotypes, from which only one haplotype is significantly associated with plant maturity. We correlated the presence of this intra-gene haplotype

with the presence of the transposon and footprint allele, but we are unsuccessful to completely tag the causative polymorphisms. In addition we defined haplotype-specific SNPs that can be used to distinguish between late and early alleles, which can be applied in marker-assisted breeding.

Materials and methods

Targeted enrichment sequencing

Targeted re-sequencing of the *StCDF1* was performed using an in-solution enrichment with RNA bait capture (Uitdewilligen et al. 2013). In short, baits were designed on the sequences of the *StCDF1* gene (disregarding the intron) ordered as a customised SureSelect bait library (Uitdewilligen et al. 2013). The bait design included also an additional set of genes, including the neighbouring genes of the *StCDF1* gene. The use of this bait library allowed to predominantly collect sequencing reads that belong to these genes, and therefore read depths of 60× – 100× were achieved. The sequenced panel comprised 83 potato varieties, representative for the global gene pool of commercial potato, both heirloom and contemporary varieties. Sequencing reads were mapped to the reference genome (V3.4, PGSC, 2011) and genotyping was performed with FreeBayes (Erikson et al. 2008). After genotyping, variants with genotype quality (GQ) of > 26 were retained and low-quality variants were discarded. More details on the composition of this dataset can be found in Uitdewilligen et al. (2013) and supplementary file S1.

Haplotype assembly

Haplotype assembly was performed after read alignment and variant calling. To facilitate haplotype assembly, only bi-allelic SNPs were retained from the two exons of *StCDF1*. Read alignment data of each sample was processed separately, and haplotypes were reconstructed using the approach described in Chapter 3. In short, first haplotypes were reconstructed using only heterozygous variants in each cultivar. Subsequently monomorphic SNPs were inserted in the haplotypes, allowing to compare haplotypes between individuals. After haplotype reconstruction in each individual, the results were aggregated and the collection of reconstructed haplotype blocks were compared between individuals.

Pedigree-informed haplotype assembly and haplotype imputation

The fragmented haplotype assembly was improved by using genetic relationships, that were present within these 83 varieties. Within this panel, 26 varieties had a parent that was genotyped. In addition for variety Markies both parents were genotyped. In Figure 1 an example of how pedigree-informed haplotype assembly is performed. Here, Ackersegen is a descendant of Hindenburg and Allerfrüheste Gelbe. Between these two varieties haplotype blocks partially overlap. Linkage between these blocks can be resolved because two alleles are shared between both varieties, in absence of double reduction. The number of options for extending linkage between blocks was decreased by discarding haplotype configurations that are not fully in agreement with this allele flow. If based on this, reconstruction of the complete haplotype with 35 SNPs was not achieved, read alignments were manually inspected for presence of sequencing fragments that uniquely link alleles of separate blocks. In most cases the complete haplotype among these 35 SNPs can then be reconstructed. For the remaining varieties for which no genetic relations are present in this dataset, only alignment information was used to link haplotype blocks and improve the length of the reconstructed haplotypes. Based on these results we reconstructed a reference library of haplotypes, and application of Poly-imputer (Chapter 6) allowed to assign combinations of these reference alleles to each genotype. The final results after imputation were used for downstream analysis.

Identification of the footprint and transposon allele in sequencing data

Previously the *StCDF1.3* allele was found to contain a transposon insertion, whereas the *StCDF1.2* allele originated by excision of the same transposon, resulting in a 7bp footprint allele, which otherwise shows 100% sequence identity. Here we identified the occurrence of this footprint in read alignment of each genotype, by pattern matching: First, the sequence of the *StCDF1.3* allele was retrieved from BAC clone RH048D11 (Genbank: AC238025.1). Within this sequence the transposon insertion was identified, and flanking sequences of the transposon insertion site were extracted and used to screen all sequencing fragments for the occurrence of the transposon (*StCDF1.3)* and/or footprint allele (*StCDF1.2*). The presence of the transposon was determined by calculating the number of *k*-mers (*k*=10) originating from the 865 bp transposon insert that is located within the *StCDF1.3* allele. A positive match for occurrence of the

transposon was defined if multiple *k*-mers (≥ 3) were present in a single read. The presence of the footprint was identified by pattern matching using flanking sequences of the insertion site, allowing for presence of a 2-10 bp insertion, using the following pattern: 'ACTAGG.{2,10}TATCAGGAAT' and reverse complement 'ATTCCTGATA.{2,10}CCTAGT'.

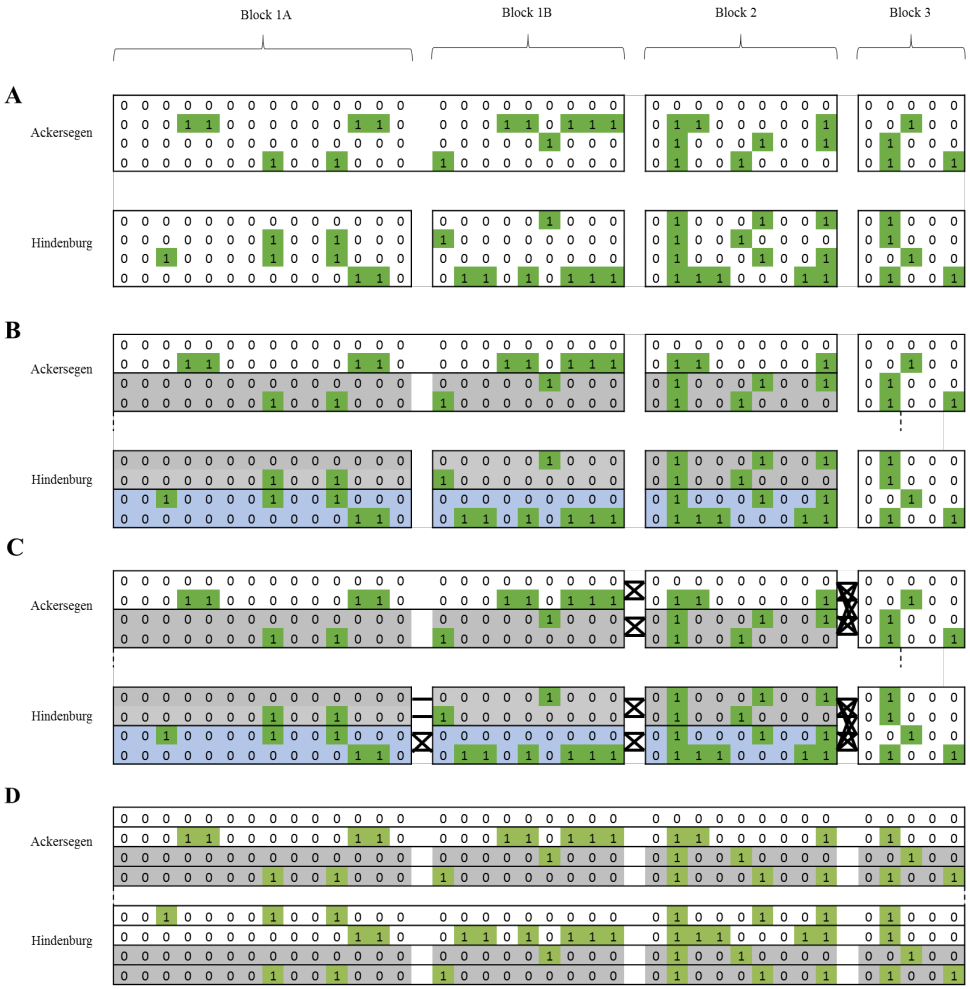


Figure 1. Example of pedigree-informed haplotype assembly variety Ackersegen (parent) and Hindenburg (offspring). **A)** The haplotype assembly resulted in a total of three disjoint blocks in Ackersegen and four disjoint blocks in Hindenburg. **B**) For block 1, 2 and 3 based on inheritance shared haplotype segments can be determined. In grey shared segments between Ackersegen and Hindenburg are visualized. **C)** These shared haplotype segments allow to diminish the potential haplotype linkages between blocks, which subsequently are solved at the level of individual reads. **D)** The resulting assembly over all 35 SNPs.

Association analysis

Phenotypic data for maturity of each of the 83 tetraploid varieties was previously described in D'hoop et al. (2008) and scored in a multi-year multi-location field trail. Plant maturity was evaluated on an ordinal scale (1=early, 9=late), based on visual field observations at the end of July-September. Adjusted phenotype means were obtained according to the description in D'hoop et al. (2011). Genome-wide Association mapping was performed in Genstat using linear regression, with including population structure as cofactor, using a strict additive model. For haplotype-based regression of intra-locus haplotypes, only linear regression was used, disregarding the effect of population-structure. In addition, multi-allelic regression was performed including all reconstructed haplotypes at the *StCDF1* gene. Subsequently, backwards selection was performed to obtain the best fit. In this procedure the least significant marker is removed until all markers are ($p < 0.05$) contributing to the final model.

Results

Haplotype analysis

A panel consisting out of 83 tetraploid potato varieties was used to identify allelic variation in the *StCDF1* gene. Previously analysis of the sequences of the *StCDF1* gene resulted in the identification of 63 sequence polymorphisms in the two exons of this gene (Uitdewilligen et al. 2012). Here we performed haplotype assembly using the short-reads (100 paired end sequencing) provided in this dataset, resulting in partial haplotype (blocks) across these exons. The first exon was not sequenced at high read depth, resulting in a large proportion of fragmented haplotype blocks. At higher read depth, a highly contiguous haplotype assembly was observed (2nd exon: PotVar0079590-PotVar0079635). Therefore, were able to obtain partial haplotype information for the 2nd exon of the *StCDF1* gene (Figure 2).

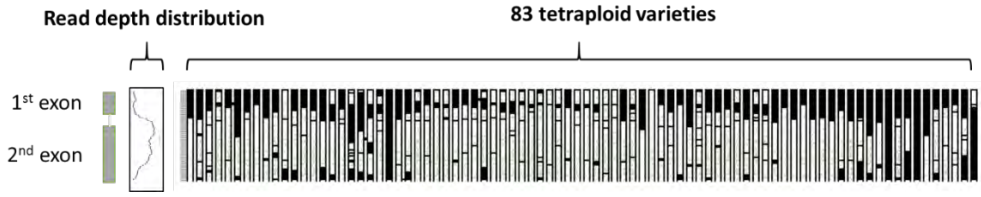


Figure 2. Haplotype assembly results in fragmented haplotype(-blocks) in the StCDF1 gene. Haplotype assembly resulted in larger blocks in the 2nd exon of the gene. The first exon of the StCDF1 gene had lower read depth, resulting in shorter haplotype blocks. In the right panel black indicates presence of smaller haplotype blocks.

After initial haplotype assembly, pedigree relations were used to manually generate longer haplotypes in 47 out of 83 varieties. If pedigree data did not allow the reconstruction of more contiguous haplotypes, read alignments were inspected if sequencing fragments uniquely linked homologous alleles within two adjacent haplotype blocks. In that case, read information was used to improve the haplotype assembly. The use of both pedigree data and read alignments resulted in 20 haplotypes across 35 SNPs, comprising 1100 bp of the 2nd exon of the *StCDF1* gene.

To obtain full-length haplotypes in the remainder of 37 varieties, which only contain fragmented haplotype blocks, we employed Poly-Imputer (Chapter 6). The application of Poly-Imputer does not discover new haplotypes, but tries to decompose genotypic data into the best combination of four, out of the 20 haplotypes. To execute Poly-Imputer we used the set of 20 haplotypes as input and performed imputation of all 83 varieties, including the 47 for which manually haplotypes were reconstructed. After imputation, 12 of the above mentioned 20 haplotypes were found in the dataset, but 8 were not observed. These eight haplotypes represent low-frequent haplotypes (1-5 counts), and likely represent haplotype reconstruction errors, that occurred during haplotype assembly or through manual processing of haplotype blocks. The reason why Poly-Imputer assigned other haplotypes to these varieties, is that a different combination of four haplotypes better explains the observed dosages of a variety, given the observed population allele frequency. After imputation, 8 out of 12 haplotypes have allele frequencies above 1% (Table 1, Figure 4), and 3 have allele frequencies below 1%. Among the haplotypes a SNP density of 1 SNP every 32 bp was found. In terms of allele composition for 22 varieties a quadrigenic (abcd) configuration was observed, whereas the numbers for trigenic (aabc), digenic-duplex (aabb), digenic simplex (aaab) and monogenic (aaaa) are 47, 10, 1 and 0, respectively. Hence most varieties contained three

unique alleles and an average number of unique alleles of 3.1. The numbers of observations per unique genotypic class ranges from 1-5, indicating that most varieties contained a unique allele composition. For the complete allele composition in each variety see File S2. To test deviations from Hardy-Weinberg equilibrium, we employed a chi-squared (χ^2) test for each allele with count > 5, by comparing expected genotypic frequencies to observed genotypic frequencies. This test showed that all reconstructed haplotypes were in close agreement of Hardy-Weinberg equilibrium.

	PotVar0079590	PotVar0079591	PotVar0079592	PotVar0079593	PotVar0079594	PotVar0079595	PotVar0079597	PotVar0079598	PotVar0079599	PotVar0079600	PotVar0079601	PotVar0079602	PotVar0079603	PotVar0079604	PotVar0079605	PotVar0079606	PotVar0079607	PotVar0079608	PotVar0079611	PotVar0079612	PotVar0079613	PotVar0079614	PotVar0079615	PotVar0079616	PotVar0079617	PotVar0079618	PotVar0079625	PotVar0079626	PotVar0079628	PotVar0079629	PotVar0079630	PotVar0079631	PotVar0079633	PotVar0079634	PotVar0079635
H1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0
H2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H3	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	1
H4	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1	0	1	1	1	0	1	1	1	0	0	1	1	0	1	0	0	0
H5	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0	1	1	0	1	1	1	0	1	1	0	0	0	0	1	0	1	0	0	0
H6	0	1	0	0	0	1	0	0	1	0	1	1	0	0	1	1	0	1	0	1	1	1	1	1	1	0	0	0	1	1	0	1	0	0	0
H7	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
H8	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	1
H9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	1	0	0	1
H13	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
H16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1
H17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Table 1. Haplotype composition in 83 varieties. The cells with black border are tag-SNPs for that specific haplotype.

Association analysis

The genome-wide association analysis identified a major-effect QTL on chromosome *5*, as previously reported in Kloosterman et al. (2013), coinciding with the presence of the *StCDF1* gene. The most significant association was found with PotVar0078096 at PGSC4.03 coordinate chr05:4408254, located 130 kb distal of the *StCDF1* gene with $-\log_{10}(P)$ of 14.98. The best association within the *StCDF1* gene was found with PotVar0079616 at PGSC4.03 coordinate chr05:4541104, with $-\log_{10}(P)$ of 13.02. These two SNPs explain 43% and 31.59% of the phenotypic variation, respectively (File S3).

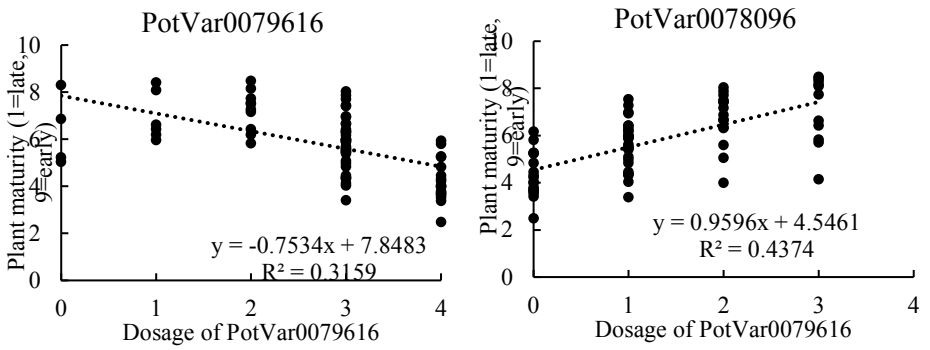


Figure 3. Correlation between the most significant SNPs (PotVar0079616 and PotVar0078096) and maturity index. Note that PotVar0079616 is positively correlated with the presence of the reference allele, whereas PotVar0078096 is negatively correlated with the presence of the reference allele.

An important issue is to compare the power of GWAS based on individual SNPs with the power of a haplotype-based GWAS analysis. To identify which haplotype has an effect on plant maturity, an association analysis was done with all 12 haplotypes. This identified a single haplotype (*H2)* that is significantly associated with $-\log_{10}(P) = 8.78$ to plant maturity, where presence of the alternative allele leads to early tuberizing varieties (Figure 4A). This haplotype contains for all SNPs the reference SNP-allele, therefore similar to the DM *S. tuberosum* group Phureja DM1-3 516 R44 StCDF1 allele. The allele frequency of haplotype *H2* is 0.36. This single allele explains 41% of the phenotypic variation. For all other 11 haplotypes, no significant association was found.

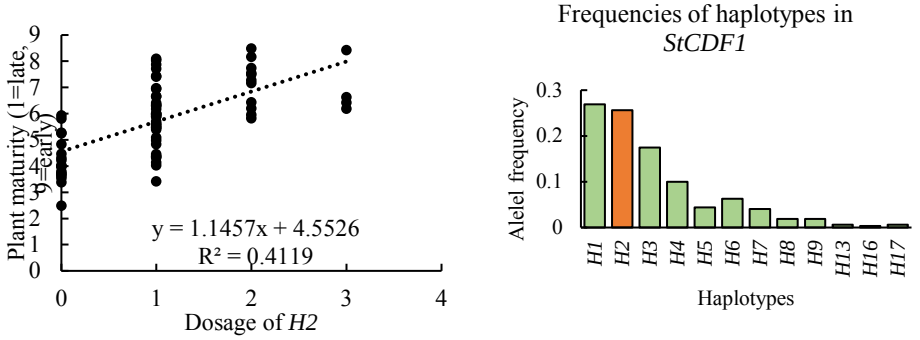


Figure 4. A) Correlation between plant maturity and dosage of haplotype H2 B) Frequencies of haplotypes observed in the StCDF1 gene. The orange haplotype leads to an early tuberization response.

Presence of transposon and/or footprint allele

Previously two mutant alleles were identified that lead to early maturity (Kloosterman et al. 2013). The *StCDF1.3* allele contains a 865 bp long transposon insertion, whereas the *StCDF1.2* allele originated through an excision event, and only contains a insertion of 7bp. Both these inserts cause a premature stop codon resulting in truncated proteins, and should lead to a similar phenotypic effect. In this study we identified the presence of either transposon or footprint with pattern matching that identified portions of the transposon or footprint in sequencing reads. From the 83 varieties, 67 varieties contained sequencing reads with the presence of a 7 bp insertion at the transposon- insertion site. Analysis of transposon-containing sequencing fragments identified 47 varieties that contain one or more copies of the transposon insertion (File S2). Here, we could not estimate discrete dosages of this allele, but we associated the proportion of reads (as

proxy for dosage), originating from either transposon or footprint as predictor for early maturity. We observed a significant association between the proportion of transposon-containing reads and plant maturity of $-\log_{10}(P) = 8.94$ (Figure 5A). Surprisingly, no association was observed between occurrence of the footprint allele and plant maturity (Figure 5B).

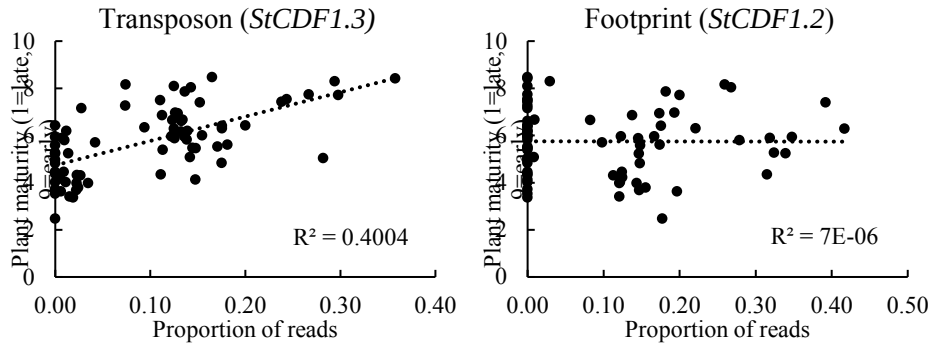


Figure 5. Correlation between plant maturity and proportion of reads originating from (A) transposon or (B) footprint allele.

Correlation analysis

To understand which of the reconstructed haplotypes contains the transposon insertion or footprint, the squared correlation coefficient (r^2) between proportion of reads per sample and dosages of a single haplotype was used. From all haplotypes, only *H2* has a strong correlation to the presence of transposon-containing reads (r^2 of 0.29), as visualized in Figure 6. In addition also haplotype *H1* (AF=0.26; $r^2 = 0.1$) showed a weak correlation with presence of the transposon, whereas all other haplotypes have no substantial correlation to the presence of the transposon. We also evaluated the correlation between presence of the proportion of reads originating of the transposon allele and all SNPs within the haplotyped region, from which we observed only for PotVar0079616 a significant correlation with r^2 of 0.26. We also evaluated the correlation to the (distal) SNP PotVar0078096, resulting in a r^2 of 0.36. In addition, the presence of the 7bp-footprint excision motif was correlated with *H2* ($r^2 = 0.13$) and haplotype H3 ($r^2 = 0.68$). The latter haplotype is not associated with plant maturity ($-\log_{10}(P) = 0.46$). Here we also observed a correlation between *H3* and *H2* ($r^2 = 0.18$), and between *H1* and *H2* ($r^2 = 0.18$).

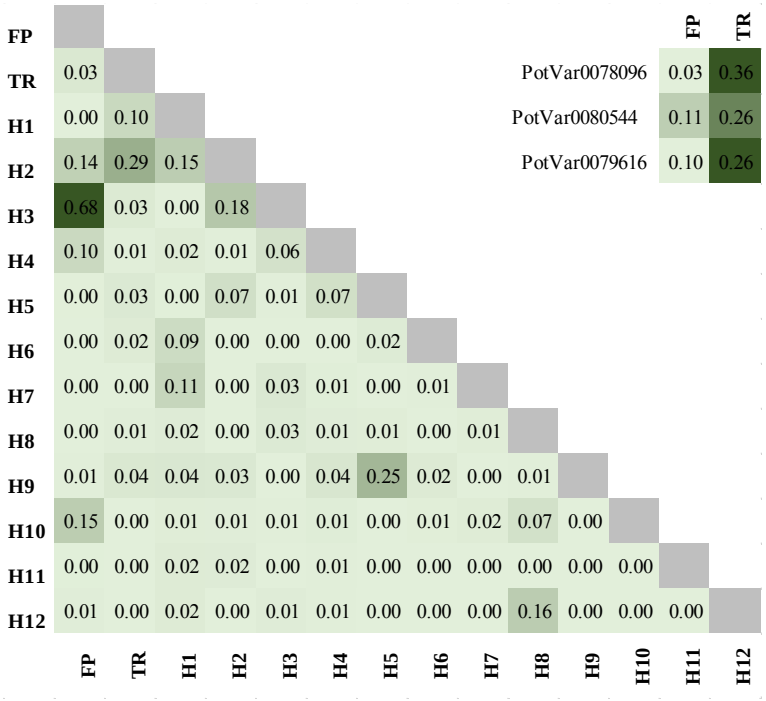


Figure 6. Correlation matrix between haplotypes and occurence of footprint/transposon. 'FP' indicates footprint occurence and 'Tr' indicates presence of transposon. In the lower triangle, the results of correlation between haplotypes, and between haplotypes and 'Tr' and 'FP'. The upper panel reflects the correlation between 'FP' and 'Tr' and individual SNPs.

To investigate the possible contribution of multiple alleles we employed multi-allelic regression using the twelve reconstructed haplotypes (*H1-H12*). The use of all haplotypes in a single regression model resulted in an explained 56% of the phenotypic variation. The application of forward and backward regression allowed to define a model with only nine haplotypes, resulting in a similar explained variation of 55.6%. This model contained *H1, H3, H4, H5, H6, H7, H8, H9, H11*, but not *H2*, *H10* and *H12*, effectively selecting against presence of *H2*.

Discussion

In this study we used next-generation sequencing to identify haplotypes within the *StCDF1* gene, a major regulator of daylight-dependent tuberization, colloquially known as plant maturity. In total, 20 haplotypes were identified in 47 tetraploid varieties for the 2nd exon of the *StCDF1* gene. After manual curation, and imputation, 12 haplotypes were

retained, whereas eight haplotypes were discarded as these likely represent erroneous assembled haplotypes. To each variety the best combination of four out of these eight haplotypes were assigned, allowing to determine the haplotype composition of all 83 varieties. Of these eight haplotypes, only one haplotype displays a significant association with plant maturity. The dosage variation of this haplotype explains 40% of the phenotypic variation in plant maturity.

Haplotype analysis

From the reconstructed haplotypes we can conclude that the *StCDF1* gene is characterized by an enormous allele diversity. Previous studies have identified haplotypes based on Sanger-sequencing for other genes such as the *PHO1a* and *StGWD1* gene, indicating that commercial potato contains between 9-16 haplotypes (Schreiber et al. 2014, Uitdewilligen et al. 2012), which is roughly similar as the amount of alleles (12) discovered here. Likewise in a previous study we estimated that on average 15 unique haplotypes are present within a segment of 25 SNPs, based on haplotype analysis of 800 genes (*Chapter 3).* Typically these haplotypes display a frequency spectrum that follows an exponential distribution, with a few common haplotypes and a larger tail of rare haplotypes. The rare haplotypes with an allele frequency below 1% might represent recent wild species introgression segments, which can be validated with pedigree information, as these haplotypes show identity-by-descent. Otherwise, the rare haplotypes are erroneous due to genotyping errors or haplotype reconstruction errors.

So far most studies that have reported on haplotype analysis in potato have made use of amplicon-sequencing, which is suitable for small scale investigations of allele diversity at a single or a few loci (Rickert et al. 2002, Schönhals et al. 2008, Schreiber et al. 2012, Uitdewilligen et al. 2012). Interpretation of amplicon sequences allows SNP calling, but reconstruction of the underlying haplotypes requires sub-cloning of the PCR product and colony sequencing. (Uitdewilligen et al. 2012). In this study we used short-read sequencing technology to collect linkage phase information between the SNP alleles as contained by the reads (read-pairs) to reconstruct haplotype diversity at a single locus. In contrast to PCR amplicon-based strategies, the length of a reconstructed haplotype is not limited by the length of the PCR amplicon, but rather by sequencing characteristics such as read depth, library insert size and read length (*Chapter 3).* A downside of this sequencing-based approach is that fragmented haplotype blocks are recovered initially,

for which no linkage between adjacent haplotype blocks is known. Improvement of these solutions requires either manual haplotype reconstruction, or the use of other sources of haplotype information. In this paper we improved the contiguity of the haplotypes by manual phasing, or pedigree relations, as well as by manual assignment of sequencing reads. After obtaining a haplotype library, imputation with Poly-Imputer (*Chapter 6*) was performed, allowing to quickly screen additional genotypes for the presence of these haplotypes.

Association analysis

Strikingly the position of the SNP marker PotVar0078096 with the most significant association with plant maturity was at 130 kb distance from the causative *StCDF1* locus. This SNP allowed to explain 43% of the phenotypic variation. In contrast, the most significant marker residing within the *StCDF1 gene* explains only 31.6% of the phenotypic variation (PotVar0079616), which indicates that this marker does not tag the causative allele unambiguously. Association with haplotype *H2* allows to explain 41% of the variation, and judged from the haplotype composition, the SNP PotVar0079616 is not completely haplotype-specific for the *H2* allele (e.g. the reference allele is present in *H16* and *H17)*, which likely explains the difference in strength of association to maturity differences between *H2* and PotVar0079616.

In our study none of the other 12 alleles within the *STCDF1* have a significant correlation to maturity, indicating that differences in plant maturity are likely to be mediated by a single allele at the *StCDF1.* The *H2* allele allowed to explain 40% of the phenotypic variation. The use of all haplotypes at this locus increased this to 55%. In view of the high broad sense heritability (H^2 = 0.85 - 0.9) of plant maturity (D'hoop et al. 2009, 2011), a considerable part of the genetic variation is not explained by this locus, suggesting an important role for epistatic and allele interactions and potentially a lack of QTL detection for other minor QTLs. Previously we identified a QTL for plant maturity on chromosome *3*, which was only detected using haplotype-based GWAS, and was not found in this study (*Chapter 5*). In this study the chromosome *3* QTL was not detected, possibly because of the small sample size (N=83). Another reason could be that single SNP markers fail to adequately tag the causative chromosome *3* allele, which subsequently can escape QTL detection (See *Chapter 5*).

Kloosterman et al. (2013) showed that multiple variants of the *StCDF1* gene could lead to early maturity. We identified the presence of transposon and/or footprint alleles in each sequenced variety. Correlation analysis show that presence of the transposon is associated (r^2= 0.29) with presence of haplotype *H2* (Figure 5), and also exhibits a weak correlation with haplotype *H1* (r^2 =0.1). If these two haplotypes would perfectly tag the presence of the transposon we would expect a correlation that is close to 1.0. Here we observe that the correlation between *H2* is much lower than expected, which likely can be explained by the lack of discrete dosage assignment of transposon presence.

Surprisingly no association was observed between early plant maturity and the presence of the footprint (*StCDF1.2)*. Based on correlation analysis the presence of the footprint is strongly correlated to presence of *H3*, and also *H2*. In a previous study an allele containing the footprint excision event displayed a strong phenotypic effect. Notably within in the C×E population (*See* S1 Kloosterman et al. 2013), clones originating from the *C×E population* containing a single copy of the *StCDF1.2* clearly show an early phenotype (Kloosterman et al. 2013). In addition complementation tests of this allele in *Solanum andigenum* and late maturing C×E progeny, both led to early tuberization (Koosterman et al. 2013). In this study we could not find an association between presence of 7bp-insertion and plant maturity, which is at odds with previous knowledge.

Is *H2* the causative allele that contains the transposon insertion?

For marker-assisted breeding haplotype-specific markers are required. So far, the best marker is the 130 kb distal of the *StCDF1* gene (PotVar0078096), as judged from the association analysis, and this marker is also correlated to the presence of the transposon. As judged from the correlation analysis the intra-gene *H2* haplotype correlates strongly to the transposon presence, and is correlated weakly with occurrence of the footprint or wild type alleles. This suggest that *H2* should be split into a $H2^-$ (without transposon) and $H2^+$ (with transposon). The presence of *H2* could be detected with a combination of PotVar0079616, PotVar0079625 and PotVar0079626 can be applied for use in marker-assisted selection, whereas at this moment no distinction can be made between $H2^-$ and $H2^+$. The separation of these two alleles can likely be achieved by either constructing longer haplotypes, or by developing a marker specifically targeting the presence of the transposon.

Conclusions

In this study we identified a single allele of the *StCDF1* gene that explains differences in plant maturity, and likely contains the transposon insertion. This haplotype does not distinguish completely between alleles containing a transposon and alleles that do not contain the transposon insertion. Therefore we postulate the presence of a *H2*$^{-}$ (without transposon) and *H2*$^{+}$ (with transposon). We fail to understand why presence of the footprint has no correlation with plant maturity. In this study we demonstrated that next-generation sequencing allows to characterise haplotype diversity of a single locus. The use of genetic relations and haplotype imputation allows to build complete haplotypes in all sequenced varieties.

Acknowledgements

JHW is supported by a grant of the Dutch Science Organisation NWO (project 831.14.002).. The potato breeding companies Averis Seeds B.V., HZPC Holland B.V., KWS POTATO B.V. and Meijer B.V. are acknowledged for contributing phenotypic data. The Dutch Technology Foundation (STW grant WPB-7926) financed the sequencing data.

Additional files

File S1: Description of the variety panel of 83 varieties

File S2: Supplementary file S2: StCDF1 reconstructed haplotypes.

File S3: Supplementary file S3: GWAS for plant maturity over 135000 variants

Chapter 8

General discussion

This thesis deals with two aspects important in potato breeding and genetics. Firstly, the development of methods and tools to identify haplotypes in tetraploid potato. Secondly, application of these methods and tools to reconstruct haplotypes and use of these haplotypes in genetic studies. In the second chapter, I started with a genome-wide association mapping study (GWAS) for potato tuber shape, which identified a major effect QTL on potato chromosome 10. Subsequent fine mapping in an experimental diploid potato population (C × E), refined this locus to a region of ~200 kb. Within this population, multiple alleles, conferring round or long tuber shape were found to segregate and jointly explained phenotypic trait variation. However, within an association panel, consisting of 537 varieties, no knowledge was available about haplotype composition, limiting our understanding of the genetic architecture of potato tuber shape. To overcome the lack of knowledge of haplotype composition, tools were developed to identify haplotypes, starting with the development of a computational method that reconstructs short-range haplotypes from short read sequencing data (Chapter 3). Subsequently, a method for long-distance haplotyping was developed, which uses SNP genotyping data (Chapter 4). This tool was applied to SNP data of a potato association panel, comprising 537 tetraploid varieties (Chapter 5). Because the results from the methods for haplotype reconstruction that were introduced in chapter 3 and 4 often are incomplete or contain errors, I explored the use of a curated library of reference haplotypes to impute those haplotypes on unphased genotypes (Chapter 6).

Once these tools were developed, I explored the use of reconstructed haplotypes in an association mapping experiment. In Chapter 6, previously identified QTLs for tuber shape (Chapter 2), maturity (Kloosterman et al. 2013) and flesh colour (Wolters et al. 2010) were re-evaluated with haplotype-based association mapping. This had the goal to link these QTLs to allelic information. Also, Chapter 7 describes an in-depth study to uncover allelic variation of the *StCDF1* gene, a key regulator of day-light dependent tuberization.

In this general discussion, I will discuss the relevance of findings that were obtained in the six experimental chapters, and try to place these in a broader scientific context. Here I will focus on the discussion of aspects that encompass haplotype detection methods, and implications of using haplotypes for QTL discovery, specifically in the context of an association mapping panel. Last but not least, I will discuss the application of haplotypes

in marker-assisted breeding, and steps that need to be taken to include knowledge about haplotypes in a practical potato breeding program.

Prologue

In science, no discovery is made in solitude, and that is certainly true for this thesis. Most of the presented findings in this thesis evaluate existing ideas. This thesis builds further on research that was performed in three previous *PhD* projects (D'hoop, 2008; Uitdewilligen, 2013; Vos, 2017), which were executed at the department Plant Breeding of Wageningen University & Research. Almost a decade ago, D'hoop et al. (2009) performed association mapping in a large set of potato varieties, using AFLP technology to identify marker-trait correlations with agronomical traits. Later, a broader panel was genotyped with the SOL-STW SNP array, comprising 15K usable SNP markers (Vos et al. 2015). Part of these SNP markers were found in a SNP discovery study in 2013, where targeted resequencing of 800 potato genes allowed to characterize 135,000 DNA sequence variants (Uitdewilligen et al. 2013). Re-sequencing showed the enormous genetic diversity of the potato genome and led to the understanding that sequencing reads could be used to perform haplotyping in potato (Chapter 3). The SNPs identified in this set of varieties allowed to construct the high-density SNP array that was previously mentioned (Vos et al. 2015) and generated insights related to agronomical traits in potato (Kloosterman et al. 2013; Vos et al. 2016). The application of association mapping using data from this SNP array led to the understanding that haplotype information is crucial for application of markers in potato breeding. In this thesis, I provide tools that allow to routinely interrogate both the sequencing dataset of Uitdewilligen (2013) and the SNP-array data (Vos et al. 2015) to obtain a catalogue of naturally occurring haplotypes, and explore the application of these haplotypes in genetic studies.

Genomics in polyploids: Development of sequencing technologies

In recent years, much progress has been made in plant genomics due to development of platforms that generate long reads and increasing sequencing capacity (Goodwin et al. 2016). Whereas in the previous decade it was only possible to reconstruct a single reference genome with a lot of efforts (PGSC, 2011; ITAG, 2009), nowadays the construction of multiple references has become feasible, even for small research groups. Examples of plants for which multiple reference genomes are available are *Arabidopsis thaliana* (Gan et al. 2011; Jiao et al. 2017) and tomato (The 100 Tomato Genome

Sequencing Consortium, 2014). The increased availability of these fully phased reference assemblies indicates that the reconstruction of such an assembly poses fewer challenges than before: For instance, a *de novo* assembly of the wild tomato species *Solanum penneli* was reported using only Nanopore sequencing data (Schmidt et al. 2017). Likewise, phased assemblies have been reported for crops such as pepper and grape (Hulse-Kemp et al. 2018; Minio et al. 2017). In polyploids, such as potato, progress has lagged behind. For instance, already in 2011, the potato genome was sequenced, using a doubled monoploid, whereas, at that time, assembly of the heterozygous diploid clone RH89-039-16 failed due to excessive heterozygosity (PGSC, 2011).

Almost seven years later, a fully phased tetraploid potato assembly has yet to be reported. For other polyploid crops such as the heterozygous hexaploid sweet potato recently a partial phased assembly was published (Yang et al. 2017). In the case of the sequencing of sweet potato, partial phasing was achieved by using a reference-based phasing procedure (Yang et al. 2017). Likewise in the tetraploid rose (*Rosa chinensis*) a reference genome was still constructed from a doubled haploid line (Hibrand et al. 2018). These reports show that sequencing highly heterozygous polyploid species is still a challenge.

A pre-requisite for the efforts to generate a fully phased assembly is the availability of efficient phase reconstruction methods. Indeed in diploids, a key development for obtaining fully phased assemblies was the development of reliable tools that allow disentangling the two alleles in highly heterozygous diploid genomes (Jiao and Schneeberger, 2017). In this thesis, a start was made with the development of such methods for polyploids. Here we developed a reference-guided haplotype assembly method (Chapter 3), as at that time this seemed the most suitable approach to explore the reconstruction of haplotypes, and because only short reads were available of a targeted resequencing effort of 83 potato varieties (Uitdewilligen et al. 2013).

A reference-guided haplotype assembly method

Accurate phase detection in polyploids can be divided into two stages. Firstly, the detection of haplotypes given a set of reads. Secondly, the determination of the dosage of these haplotypes. In Chapter 3, I introduced an approach that allows achieving these goals simultaneously, using high quality paired-end short-read sequences, originating from Illumina sequencers. Phase reconstruction was achieved with a probabilistic model that models sequencing errors and haplotype dosages. At reasonable read depth (> 40×)

highly accurate short haplotype blocks were reconstructed, suitable for downstream application in genetic studies.

Discontinuity of phasing solutions

One of the insights I obtained during development of the haplotype assembly method is that the output of haplotype assembly, as performed within a single individual, is not easily comparable between different genotypes. Phasing often results in a set of discontinuous haplotype blocks that are only partly overlapping between individuals (Chapter 3, Figure 7; Chapter 7, Figure 2). A single region is then split into many discontinuous haplotype blocks. Indeed, sequence polymorphisms are not randomly placed across the four homologous chromosomes, and also not uniformly distributed across genes or chromosomes. As a result, these haplotype blocks are only fully haplotyped in a subset of all genotypes, which poses a challenge to how to apply these blocks in for example association mapping.

Using identity-by-descent to improve accuracy and completeness of haplotype solutions.

A potential strategy to cope with this fragmentation, apart from increasing the quality of the sequencing data, is to use genetic relations that occur between individuals (pedigree relations). These relations can be used to improve phase estimation, as the assumption of identity-by-descent (IBD) implies that individuals share multiple haplotypes. The advantage of incorporating this information to estimate haplotypes can be seen in a recently proposed method uses IBD to phase a polyploid father-mother-child trio (Motazedi et al. 2017). Provided that both parents and progeny are selected for sequencing, this allows to improve phasing continuity, but at the expense of needing to genotype multiple related individuals. In Chapter 7 an example can be seen of this. Initially, a highly fragmented haplotype assembly was obtained. The use of genetic relations allowed to improve this assembly to achieve full-length phasing over a region of 1100 bp of the *StCDF1* gene.

Limits for haplotype assembly in polyploid crops

In general, the length of sequencing-based haplotypes is limited by the length of the used sequencing reads, or in other words, the maximum distance between variants, that is spanned by a set of paired-end sequencing fragments. Only if enough sequencing reads span between SNPs, accurate phase estimation can be achieved. As a consequence haplotype assembly is not very well suited for crops with low SNP density, and more

suitable for highly heterozygous crops such as potato. In Chapter 2, I used a sequencing dataset with a limited insert size (~270 bp) and short read length (100 bp), obviously not ideal to obtain long haplotypes. Several strategies are available to improve the length of these haplotypes or contiguity of the haplotype assembly. For instance, longer haplotypes can probably be reconstructed by increasing the read length, and at the same time, increasing the fragment insert size distribution. However given that conventional Illumina sequencing can only accept up to 1 kb of inserts, it will be a big challenge to obtain gene-scale haplotypes. An alternative strategy to achieve a similar goal is by using a cocktail of short and long reads. In that case, short reads can provide short-range phase information, whereas long reads will allow the reconstruction of bigger haplotypes.

Long reads will allow the reconstruction of longer haplotypes

At the beginning of my PhD, I never envisioned the rise of third generation long read sequencing technologies. A few years ago, PacBio sequencing could already achieve a maximum length of 20kb with high error rates but only allowed low-throughput sequencing. The new kid on the block, Oxford nanopore sequencing was only in the first stage of development and generated extremely long sequencing reads with base accuracy ranging from 60-80%. Currently, both sequencing technologies allow to generate vast amounts of reasonably high quality (95-98% base accuracy) long-read (> 100 kb) sequencing data.

For sequencing data originating from these technologies, a more suitable approach for haplotype reconstruction might be to first detect all unique (partial) haplotypes, which requires less computational resources, but more importantly, less read information, followed by the quantification of the dosage of each haplotype. The application of an equivalent approach has yielded completely phased Arabidopsis genomes (Chin et al. 2016) with a combination of the FALCON genome assembler, followed with phase detection with FALCON-unzip. In short, this approach creates an assembly graph allowing to assemble contigs (FALCON), followed by phasing of heterozygous variants, and for each separate allele a so-called 'haplotig' is generated (FALCON-unzip). An advantage of these approaches is that not only single nucleotide polymorphisms (SNPs) are phased, but also structural re-arrangements, insertions and deletions are processed during phasing. As copy-number variations occur with high frequency in potato (Pham et al. 2017), the need for such approaches is obvious.

Phasing in absence of physical proximity (statistical phasing).

Long-distance phasing typically is not achieved by using sequencing reads, but rather with statistical estimation of linkage phase between SNPs. In chapter 4, I developed an approach which allows to achieve long-range phasing, using population-wide genotypic data. The usage of this haplotype reconstruction method allowed to generate haplotypes for almost 15K markers, with good predictability for haplotypes that have high allele frequency, but lacks resolution for the reconstruction of alleles with low frequency (i.e. low frequent alleles are often erroneous). The downside of statistical phasing is that if the population is highly skewed in terms of population structure, haplotype phasing will likely result in erroneous haplotypes. This implies that great care is required for the selection of an adequate variety panel to assess haplotype diversity using statistical phasing.

Hybrid approach for haplotype reconstruction

The results of Chapter 3 (haplotype assembly) and Chapter 4 (haplotype inference) indicated to me that both strategies for haplotype reconstruction have their own merit. Single individual haplotype assembly results in short haplotype blocks but can identify alleles with low frequency with equal confidence as haplotypes that have a high allele frequency. Statistical phasing allows to generate long-distance linkage phases in absence of sequencing data. To make use of advantages of both methods, a good strategy may be to develop a hybrid approach that uses sequencing reads to support the presence of haplotypes within an individual. In case of lack of sequencing support, population-wide haplotype phasing allows to infer phases. Such an approach will likely result in haplotypes that are reconstructed with greater confidence, and result in a more complete haplotype: Commonly occurring alleles will be reconstructed with greater confidence, as they are seen multiple times in a population. Low-frequent alleles can be verified, because of support in sequencing fragments.

A set of known reference haplotypes allow quick screening of allele diversity

In Chapter 3 and Chapter 4, I developed a method that allows to reconstruct haplotypes in tetraploid potato varieties and a start was made in obtaining a catalogue of haplotypes present in the potato gene pool. Once this was achieved, I subsequently asked a different question: Given a set of known haplotypes, how can we identify which of these haplotypes are present in other unphased genotypes? Indeed, if most or all haplotypes and their frequencies are known, the identification of haplotypes in un-phased (new)

genotypic data is trivial. Previous reports have used individual tagSNPs or haplotype-specific SNPs to identify haplotypes (Johnson et al. 2001). An example of this can be found in Uitdewilligen et al. (2012) where haplotypes in parts of the *StGWD1* gene were identified with the use of conventional Sanger sequencing. After building an initial set of haplotypes (i.e. sub-cloning of PCR amplicons), each subsequent variety was interrogated for the presence of haplotype-specific SNPs. Presence or absence of these haplotype-specific SNPs, allowed to resolve the allelic configuration of other samples.

In Chapter 5, therefore, I introduced an approach that allows to use existing haplotype data as input to impute haplotypes on unphased genotypic data. This procedure works by determining if a haplotype configuration is consistent with un-phased SNP calls. A very logical application of this approach is to use this in a bi-parental population. After linkage mapping, order and location of markers is known and each SNP allele is assigned to its homologous chromosome, and haplotype information can be extracted. Each progeny should contain a combination of these parental homologous chromosomes. Application of Poly-Imputer allowed to score these haplotypes in the progeny with an accuracy of 98%. In this case, imputed haplotypes were compared to haplotypes reconstructed with tetraOrigin, which models recombination as well. A disadvantage of the current implementation of Poly-Imputer is the requirement of segments for which limited or no recombination is observed.

Another application of this tool was to improve the contiguity of haplotype solutions obtained by haplotype assembly (Chapter 3). As mentioned above, the haplotype assemblies originating from sequencing data are highly fragmented. In some individuals, reasonable length haplotype blocks are found. The use of these blocks as reference haplotypes allowed to impute haplotypes on the remainder of the varieties.

What length of haplotypes is required?

Arguably for most applications in plant breeding phasing does not need to be on a chromosome-level and even fragmented solutions, such as obtained in Chapter 3 may be sufficient. Here we define a haplotype as: 'a segment of multiple adjacent SNPs that are present in only one homologous chromosome'. Commonly an allele can be defined as 'a variant of a gene or locus'. Hence the word 'haplotype' is used in a more comprehensive way than the word 'allele'. Theoretically, the boundaries of a haplotype can be defined as a function of historical recombination events. This is commonly done in human genetics, where low-recombining haplotype blocks were defined on the basis of patterns of linkage

disequilibrium (Gabriel et al. 2003; Wall and Pritchard, 2003). Obviously, such estimates cannot be directly applied in polyploids, because estimates for linkage disequilibrium are notoriously difficult to interpret if marker-alleles are in repulsion phase (Vos et al. 2017). If recombination occurs between two loci, the linkage between loci decreases (linkage decay), and multiple haplotypes will occur in the population. However, in potato it was estimated on patterns of LD decay, that on average 6-10 unique founder alleles are present, coupled with a haplotype block length of 2-4 Mb (Vos et al. 2016). These results indicate that to obtain fully informative markers (i.e. detecting all segments of founder haplotypes), only a limited number of SNPs might be needed. Arguably some loci might show extreme nucleotide diversity, and other loci will show increased levels of homozygosity. As a result of this, the minimum requirement of SNPs to obtain a comprehensive view of haplotype variation at a locus is very difficult to assess.

Haplotype diversity in potato

The gene pool of commercial potato is characterized by a high nucleotide diversity, arranged in a limited set of haplotypes. In Chapter 3 we interrogated > 800 gene regions and identified haplotypes in 83 potato varieties. From these results it was concluded that on average 13 haplotypes are observed in this population, considering a segment of 15 SNPs. However, if this window was increased to 25 SNPs, on average 15 haplotypes were observed. Clearly, a pattern of diminishing returns is observed, which is in line with the above-mentioned suggestion that for accurate assessment of haplotype diversity within a potato genotype panel there is no need to haplotype hundreds of SNPs, but a moderate number (25-50) would be sufficient to capture most of the (common) allelic diversity at a single locus. On the contrary, to identify all alleles that are low-frequent, a larger number of SNPs might be needed.

The extremely high allele diversity as observed in this thesis is in line with previous results in potato, where the number of haplotypes in a part of a gene ranges between 5 and 20 (Uitdewilligen et al. 2012; Wolters et al. 2010). Each individual variety seemingly contains 3.1 unique alleles. Likewise, we observed an exponential decrease of allele frequencies, where only a limited set of haplotypes with moderate to high frequency, cumulatively explain most of the allelic variation, and a large amount of rare haplotypes that have low allele frequencies. This, is also expected given the high inter-connectively of the potato gene pool (van Berloo et al. 2013), where each potato variety is likely to be

composed of a mosaic of common haplotypes. The large amount of rare alleles might be explained by recent introductions of introgression segments, or be due to phasing errors.

Building a haplotype map of tetraploid potato

In this thesis, we started with the development of tools to assess haplotype diversity in potato. The resulting haplotype data was fragmented (Chapter 3), but does provide a good overview of haplotype diversity in potato. This provides an excellent starting point to start building a reference collection of haplotypes, especially if large amounts of sequencing information of many varieties will become available in the near future.

To succeed in building a complete haplotype map of potato I propose the following strategy:

1. Select a comprehensive set of modern varieties, accompanied by important founders (i.e. resistance donors, progenitors, landraces) and perform high depth re-sequencing.
2. Perform haplotype assembly within regions of high-quality sequencing data, and perform long-distance phasing, using a haplotype inference method.
3. Subsequently, sequence a large collection of other genotypes with low coverage.
4. The reconstructed haplotypes of step 2 can be imputed on the low-read depth re-sequenced varieties.

If such a 'core' collection is large enough, it likely covers most of the allelic variation present in most commercial potato (i.e. common alleles). Certainly, such a collection will not contain all naturally occurring alleles (i.e. rare alleles), hence new haplotype variation might be added in a later stage to the catalogue if needed. This resource will allow to quickly characterise the allelic diversity of a candidate gene or find haplotype-specific markers for application in marker-assisted selection.

The use of a catalogue of alleles in genetics and potato breeding

Expression of genes

From a molecular perspective phase information is necessary to understand how cis-and trans-acting variants can affect the expression of genes. Reports in human genetics have shown that binding affinities of different homologous alleles result in substantial variation in expression (Kasowaki et al. 2010). There is no reason to assume that this will not the case in the highly heterozygous potato. Indeed, a recent study showed that allele-specific expression is widespread in potato (Pham et al. 2016), and a large number of

genes display allele-specific expression. The results of that study were based on single nucleotide polymorphisms, but only with phase information it can be fully determined which allele is preferentially expressed.

Allele-dosage effects

Looking at studies in diploid crops such as tomato, we clearly see examples of allele-dosage effects. For example in tomato fruit size is regulated partly by an allele of the *fw2.2* gene, where more copies lead to bigger tomato fruit size (Frary et al. 2000). This phenomenon of dosage-dependent phenotypic effects is likely to be a key player in control of quantitative traits, as suggested by Birchler et al. (2001). In this thesis, we observed these effects for traits such as tuber shape (Chapter 2 and Chapter 5) and plant maturity (Chapter 5). For instance, the responsible allele for earliness, *StCDF1.2,* was long-time thought to exhibit dominant gene action. Indeed, given the molecular action of the *StCDF1,* we might expect that lack of degradation of the protein should display a dominant effect (Kloosterman et al. 2013). However, in our analysis, this allele is found to exhibit incomplete dominance (Chapter 4, Figure 5*).* Another example of a dosage-phenotype relation in potato was previously found for the *StGWD1* gene, involved in starch phosphorylation (Uitdewilligen et al. 2012), but also other genes such as *GBSS*, involved in the production of amylose (van der Wal et al. 2001; Flipse et al. 1996). In these studies, dosage-effects were observed for GBSS activity, but strikingly, the increased activity of GBSS did not result in a linear increase in amylose content, but rather the activity of GBSS was not limited anymore for the production of amylose. These examples suggest that dosage-phenotype interactions, such as the ones described above are common in potato.

Multiple alleles

A central theme throughout this thesis is the occurrence of multiple alleles and their joint effect on plant phenotypes. Most of these questions still remain. If these multiple alleles would occur, what would their contribution be to quantitative trait variation? How widespread is multi-allelism in reality? And can we disentangle the effects of these alleles? Such discussion is only warranted if we look back in history: In 1918, Fisher proposed to use the infinitesimal model for trait variance and postulated that genetic variance in a population was strictly due to a large number of Mendelian factors, each with a small additive contribution to their respective traits (Fisher 1918). During the previous century, this model has become the basis for many applications in genetics and breeding.

Nonetheless, in recent years others have questioned parts of this model, as it might represent a simplified view of the biological reality (Orr et al. 2005; Nelson et al. 2013; Hill et al. 2009), whereas aspects such as missing heritability, allelic heterogeneity, epigenetics, and occurrence of variance quantitative trait loci (vQTLs) are known to influence trait variation, but are neglected with a strict additive modelling of trait variation.

Strikingly this reservation with using the infinitesimal model as assumption, was already addressed in the early 20th century. In 1927 Sirks postulated that multiple 'allelomorphs' at a single locus could also lead to quantitative trait variation as opposed to multiple factors. For instance potato tuber shape is modulated by a major effect QTL at the *Ro* locus, which displays multiple alleles (van Eck et al. 1994; Chapter 2*).* Likewise from a theoretical standpoint a combination of different alleles at a single locus can lead many genotypic classes, and cause a wide range of phenotypic variability (Figure 1). For instance, a single locus with three functional alleles results in 9 phenotypic classes, which depending on the genetic contribution to the trait, will result in a nearly continuous phenotype distribution(Figure 1C).

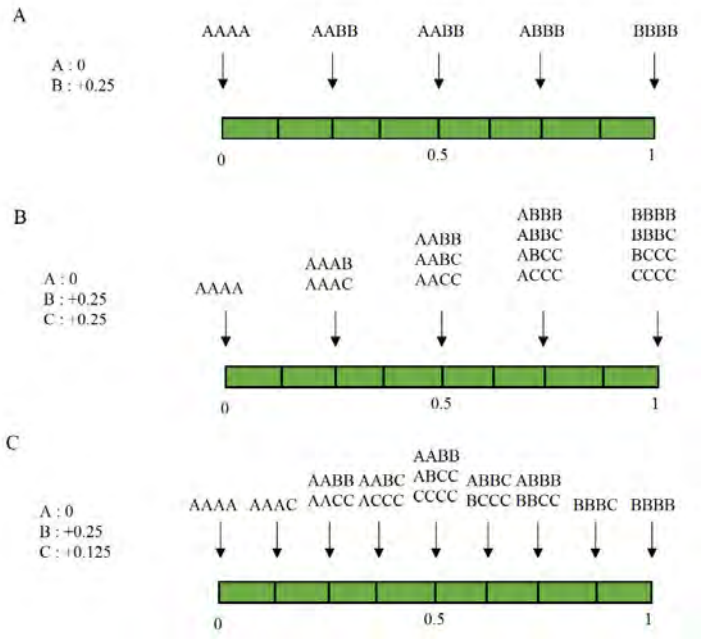


Figure 1. Expected phenotype classes for several genetic scenarios of multiple allelism. **A)** Two alleles **B)** Three alleles with distinct effects **C)** Three alleles, where B and A have equal effect.

Therefore, the idea of multiple allelism is still as relevant as it was almost 90 years ago. Especially in polyploids with a high allelic diversity, combinations of distinct alleles at a limited set of loci should suffice to generate continuous phenotypic variation. When looking at quantitative trait variation, the tendency to immediately jump to the conclusion of a complex genetic architecture, might come back in a later stage as a boomerang. While multiple alleles are clearly difficult to detect (as can be seen in this thesis), a combination of a few loci with multiple functional alleles that have different effects, will be able to explain most or all trait variation.

Knowledge of QTN(s) is needed to assess contributions of individual alleles

To assess the contributions of each individual allele on phenotypes, and show if these are caused by different causal polymorphisms, knowledge is needed about the quantitative trait nucleotide (QTNs). Indeed series of functional alleles could lead to a multitude of phenotypic classes (Figure 1C), showing quantitative trait variation. Only with haplotypes, sets of alleles sharing the same QTL can be defined, and their molecular action can be verified. But even with using haplotypes, a different causal polymorphism might be present in one or multiple haplotypes, diminishing the extra discriminatory power of haplotypes versus single SNP markers.

Multiple alleles and plant maturity

A similar situation is observed for plant maturity where haplotype-based GWAS found a single haplotype in the region surrounding the *StCDF1* gene. Previously this QTL was known to be mediated by multiple alleles (Kloosterman et al. 2013). Here we could not verify that observation which was done in a full-sib diploid population. In fact, using a small genotype panel in Chapter 7, no association was found between the presence of the *StCDF1.2* allele and maturity index, and only the occurrence of the transposon showed a strong association. Based on the results as presented in this thesis, we cannot conclude that multiple alleles have significant effects, nor exclude the possibility that multiple functional variants are present for this gene. The use of only haplotypes did not reveal this, as footprint allele and transposon allele are both likely tagged by the reconstructed haplotype.

Multiple alleles and tuber shape

In Chapter 2 it was observed that the major-effect QTL on chromosome 10 (Chapter 2) exhibits multiple functional alleles. Within the C×E diploid population only one allele conferring an elongating effect (*ro)*, and two round alleles with slightly different effects were found ($ro_{♂}$, $ro_{♀}$). For the same QTL, the results of the association mapping indicated that only one has an elongating effect (*Ro1*), but three alleles were identified with rounding effects (*Ro2-4*). If any of these alleles, conferring roundness is present, the phenotypes are similar. From a molecular perspective, the putative round alleles may represent similar (functional) alleles, whereas the elongating allele could represent a null-mutation. This finding was done in an association panel, and the causative region as defined by fine mapping is 300 kb downstream of the location of these reconstructed haplotypes. Given the extend of linkage decay it is very likely that recombination between the causative mutation and the haplotype locus could have taken place, seemingly resulting in new alleles, whereas they have a similar molecular mechanism.

For this major-effect QTL for tuber shape multiple scenarios are possible. Firstly, the three alleles confering roundness as found in Chapter 5 are recombinant alleles, as the region in which the *Ro* locus is found is located approximately 300 kb downstream of the markers used in association mapping (Figure 2B). Given the linkage decay in potato (Vos et al. 2017), it is likely that recombination has occurred between these SNPs and the causative gene. Such recombinant alleles intuitively point towards multiple allelism (Figure 2A), but just as likely this could point towards the occurrence of recombinant alleles (Figure 2B). Hence, lack of knowledge of functional polymorphisms hinders the interpretation of these haplotypes.

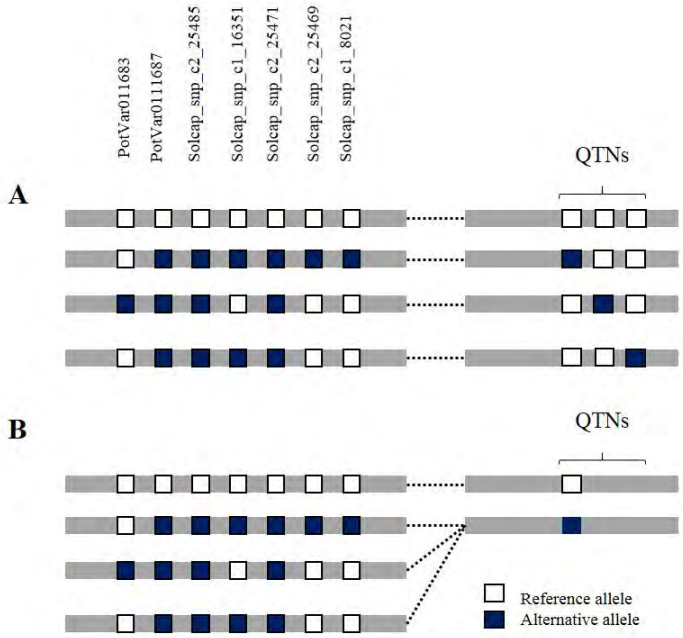


Figure 2. The haplotype structure of the *Ro* locus. Two scenarios are possible. **A)** Three functional alleles conferring round tubers. **B)** Due to linkage decay the causative alleles has recombined, and only a single causative polymorphism is responsible for differences between round and long tubers.

Such observations have been made previously in the context of association mapping in *Arabidopsis*, where conclusive evidence that multiple alleles at the *FRIGIDA* gene, involved in flowering, was only found after positional cloning of the gene, followed by the identification of nine different loss-of-function mutations as basis for differences in flowering time (Gazzani et al. 2003). Later analysis by Atwell et al. (2010), followed a similar approach as done in Chapter 2 and Chapter 6, and used association mapping to detect presence of multiple alleles by cofactor analysis, but it could not be demonstrated that they were due to the presence of multiple (dis)-functional alleles at the *FRI* gene. Likewise, association mapping for sodium accumulation in Arabidopsis identified a strong association at the *AtHKT1* gene, but multiple SNPs in this region had a strong association, suggesting either allelic heterogeneity or the presence of one or more untyped variants that are partially represented by the used SNPs (Segura et al. 2012). Also, in this case, an earlier study identified multiple alleles influencing sodium accumulation (Baxter et al. 2010).

In conclusion, knowledge of haplotypes in LD with QTN(s) allows the identification of these QTN(s) and their effects. In the case of potato tuber shape, to investigate the functional effects of the *Ro*-locus, further fine mapping and/or positional cloning is needed to demonstrate which gene is responsible for variation in shape. Subsequent investigation of allelic variants, followed by complementation tests will allow to disentangle the effect of each allele.

Towards Marker-assisted breeding

For PhD students working in potato genetics is it a tradition to include a part in the discussion about the possible application of markers in marker-assisted breeding (van Eck et al. 1995; Uitdewilligen et al. 2013; D'hoop et al. 2013; Vos et al. 2017). Obviously, the success of the application of marker-assisted selection can only be judged from the goal of the breeding program. If breeders have as goal to select for resistance genes, maturity or tuber shape, such goals are relatively easy to achieve. But what if one was to use markers for breeding of highly polygenic traits? An example of a highly polygenic trait is uniformity (Chapter 5), for which it is non-trivial to detect high confidence-QTLs, as a many minor effect QTLs influence trait variation. Each of these QTLs may need to be validated before use in marker-assisted selection.

Haplotype-specificity is key to understanding allele-phenotype associations

An aspect of haplotype reconstruction is haplotype-specificity of individual SNPs. Previously it was suggested by Vos et al. (2017) that knowledge of haplotype-specificity of each SNP is needed before application of marker-assisted breeding in potato. As correctly concluded by Vos et al. (2017) any SNP that is not haplotype-specific, might lead to irreproducibility of GWAS results. In some cases haplotype-specific SNPs are found easily (*e.g.* see results for flesh colour, maturity in Chapter 5). In other cases it is more difficult to determine the haplotype-specificity of each SNP. The determination of haplotype-specificity requires the reconstruction of haplotypes, which by application of the approaches described in Chapter 2 and Chapter 3, can be achieved routinely. From these haplotypes, a subset of haplotype-defining SNPs can be selected, and used to screen breeding material for occurrence of (a) haplotype (s), and perform marker-assisted selection. From this thesis it clearly can be seen that information about haplotype-specificity improves the understanding of allele-phenotype associations. Generally, SNPs that are less haplotype-specific exhibit a weaker association than more haplotype-specific SNPs (as shown in Chapter 5). A haplotype marker will improve haplotype-

specificity and therefore will de-convolute an association signal to a 'real' allele-phenotype relation. Nonetheless, the improved correlation between the causative allele and marker, does not automatically imply that a haplotype-specific marker will improve prediction of traits. In case of minor-effect QTLs, the contribution of a single marker is limited, and as a result the effect of using a haplotype-specific marker will be limited as well.

Genomic selection seems a suitable choice for polygenic traits.

Genomic selection is an implementation of marker-assisted selection in which genome-wide marker data is used to select progeny and/or parents, based on the joint contribution of all markers to trait variation. The reasoning behind genomic prediction is that all QTLs segregating in a population are in linkage disequilibrium with at least one marker (Meuwissen et al. 2007; Goddard et al. 2003), implying that all markers can be used to predict trait values. In contrast, marker-assisted selection makes use of markers of only a subset of markers that are previously found to be associated with the phenotype. For marker-assisted selection haplotype-specificity is important, but for genomic prediction, it is likely that if genome-wide marker data is used, most or all alleles are captured adequately by joint presence of multiple SNPs. In that case, it is likely that the use of haplotype-specific markers will improve this as well, but improvements are likely to marginal.

Studies in potato applying genomic prediction, show that traits with simple genetic inheritance, such as maturity, have good prediction accuracies of 0.77 (Slater et al. 2014; Slater et al. 2016), whereas complex traits such as breeders visual preference, yield and boiling colour show lower prediction accuracies of respectively 0.33, 0.19 and 0.44. Other reports applied genomic prediction in the same manner, but included previously obtained significant marker-trait correlations as cofactor in their statistical model, increasing the prediction accuracy in many cases (Stich & Ingelandt, 2018). One of the reasons for this improvement might be that most genomic prediction models assume that most markers have small phenotypic effects. This is likely to be a simplification, as most traits exhibit a large number of QTLs with small effect, but also a small to moderate number of major-effect QTLs influencing trait variation (Bernardo et al. 2014). In case of inclusion of these major-effect QTLs, haplotype-specificity of markers (that are included in the model), is still of importance.

Phenotyping might be just as important as the availability of a "hapmap"

Successful application of marker-assisted selection does not only depend on having haplotype-specific SNPs but potentially more on the amount and resolution of phenotypic data (Fiorani & Schurr, 2013), which is outside the scope of this thesis. For some, this is considered the next frontier, where for instance, the application of marker-assisted selection to a polygenic trait such uniformity would require to quantify all subcomponents of uniformity, that disentangle the effects of trait heterogeneity. Likewise, subtle differences between allelic combinations, present at major effect QTLs for tuber shape and plant maturity might only be found if phenotyping can measure these differences.

Marker-assisted selection in potato: How to proceed?

In view of a practical potato breeding program, marker-assisted selection can be done at progeny level or at parent level. The latter seems more efficient, as it requires less extensive phenotyping, and can avoid costly genotyping at the progeny level. A breeder can select which beneficial trait alleles are preferred, and gradually enrich the existing gene pool for desired traits. However the usability of haplotypes versus single SNP markers could easily be overestimated. A reliance on notions of 'genetic' gain through statistical models, could simplify the complexity of a set of interconnected traits, and lead to poor decision-making in a breeding program. In my opinion, marker-assisted selection will require a catalogue of natural occurring allelic variation (Figure 3), coupled with a catalogue of QTL (-alleles). This can help a plant breeder to select parents for crossing. Genotyping needs to be performed on a wide gene pool, using either high-density SNP arrays or whole genome sequencing, to allow the reconstruction of almost all alleles, that segregate in a population. After detection of QTLs, haplotype-specific markers for each associated allele can be developed, facilitating marker-assisted breeding.

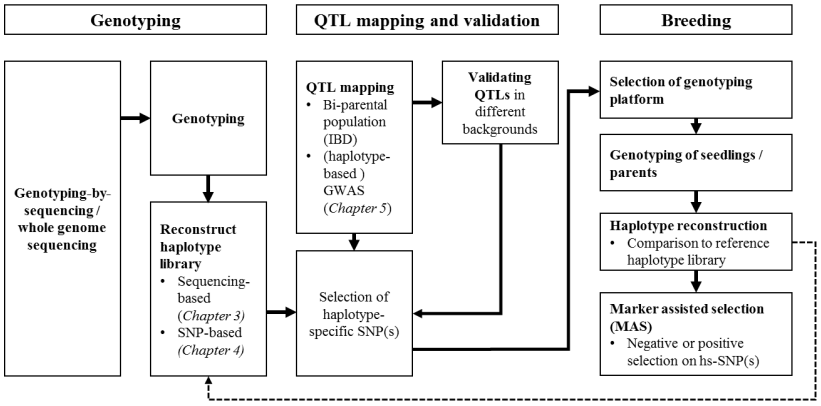


Figure 3. Schematic overview of application of haplotypes in marker-assisted selection.

Concluding remarks

This thesis has made contributions to the development of methods to identify haplotypes in the heterozygous autotetraploid potato. The reconstructed haplotypes improved our understanding of the haplotype composition of tetraploid potato. The haplotype resources generated in this study will likely result in the improvement of application of markers in practical potato breeding. The investigation of allelic diversity in potato has just started, and hopefully the methods and insights as presented in this thesis will contribute to this.

References

Aguiar D, Istrail S (2012) HapCompass: A Fast Cycle Basis Algorithm for Accurate Haplotype Assembly of Sequence Data. Journal of Computational Biology : A Journal of Computational Molecular Cell Biology 19 (6): 577–90.

Aguiar, D, Istrail S (2013) Haplotype Assembly in Polyploid Genomes and Identical by Descent Shared Tracts. Bioinformatics (Oxford, England) 29 (13).

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? European Journal of Human Genetics, 9(4), 291.

Alonso-Blanco C, Aarts MGM, Bentsink L, Keurentjes JJB, Reymond M, Vreugdenhil D, Koornneef M (2009) What Has Natural Variation Taught Us about Plant Development, Physiology, and Adaptation? The Plant cell 21(7): 1877–96.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome biology, 11(10).*

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, ... , Nordberg M (2010) Genome-Wide Association Study of 107 Phenotypes in Arabidopsis Thaliana Inbred Lines. Nature, 465(7298), 627.

Bansal V, Bafna V (2008) HapCUT: An Efficient and Accurate Algorithm for the Haplotype Assembly Problem. Bioinformatics 24 (16): 153–59

Bastien M., Boudhrioua C, Fortin G, Belzile F (2018). Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. Genome, (999), 1-8.

Baxter I, Brazelton JN, Yu D, Huang YS, Lahner B, Yakubova E, Vitek O (2010) A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1;1. *PLoS genetics, 6(11), e1001193.*

Bekele WA, Wight CP, Chao S, Howarth CJ, Tinker NA (2018) Haplotype based genotyping-by-sequencing in oat genome research. Plant biotechnology journal.

Bergelson J, Roux F (2010) Towards Identifying Genes Underlying Ecologically Relevant Traits in Arabidopsis Thaliana. Nature Reviews. Genetics 11(12).

Berger E, Yorukoglu D, Peng J, Berger B (2014) Haptree: A novel bayesian framework for single individual polyplotyping using ngs data. PLoS computational biology, 10(3), e1003502.

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. *Crop Science, 47(3), 1082-1090.*

Birchler JA, Bhadra U, Bhadra MP, Auger DL (2001) Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Developmental biology*, *234*(2), 275-288.

Bourke PM (2018) Genetic mapping in polyploids. ISBN: 978-94-6343-846-9.

Bourke PM, van Geest G, Voorrips RE, Jansen J, Kranenburg T, Shahin A, Visser RGF, Arens P, Smulders MJM, Maliepaard C (2018). polymapR: linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. Bioinformatics, 2018, 1–7

Bourke PM, Voorrips RE, Kranenburg T, Jansen J, Visser RGF, Maliepaard C (2016) Integrating Haplotype-Specific Linkage Maps in Tetraploid Species Using SNP Markers. Theoretical and Applied Genetics 129 (11).

Brigneti G, Garcia-Mas J, Baulcombe DC (1997). Molecular mapping of the potato virus Y resistance gene Rysto in potato. Theoretical and Applied Genetics 94:198-202

Browning SR, Browning BL (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics, 81(5), 1084-1097.

Browning SR, Browning BL (2011). Haplotype phasing: existing methods and new developments. Nature Reviews Genetics, 12(10), 703.

Bryan G, McLean K, Bradshaw J, De Jong W, Pillips M, Castelli L, Waugh R (2002). Mapping QTLs for resistance to the cyst nematode Globodera pallida derived from the wild potato species Solanum vernei. Theoretical and Applied Genetics 105:68-77

Buntjer JB, Sørensen AP, Peleman JD (2005) Haplotype diversity: The link between statistical and biological association. Trends in Plant Science 10(10):466-471.

Campbell R, Ducreux JML, Morris WL, Morris JA, Suttle JC, Ramsay G, Bryan GJ, Hedley PE, Taylor MA (2010) The Metabolic and Developmental Roles of Carotenoid Cleavage Dioxygenase4 from Potato. Plant Physiology, 154(2), 656-664.

Carley CAS, Coombs JJ, Douches DS, Bethke PC, Palta JP, Novy RG, Endelman JB (2017). Automated tetraploid genotype calling by hierarchical clustering. Theoretical and applied genetics, 130(4):717-726.

Carpenter, MA, Joyce, NI, Genet, RA, Cooper, RD, Murray, SR, Noble, AD, and GM Timmerman-Vaughan (2015). Starch phosphorylation in potato tubers is influenced by allelic variation in the genes encoding glucan water dikinase, starch branching enzymes I and II, and starch synthase III. Frontiers in plant science, 6, 143.

Cheng F, Sun R, Hou X, Zheng H, Zhang F, Zhang Y, Liu B, Liang J, Zhuang M, Liu Y, Liu D, Wang X, Li P, Liu Y, Lin K, Bucher J, Zhang N, Wang Y, Wang H, Deng Y, Liao Y, Wei K, Zhang X, Fu L, Hu Y, Liu J, Cai C, Zhang S, Zhang S, Li F, Zhang H, Zhang J, Guo N, Liu Z, Liu J, Sun C, Ma Y, Zhang H, Cui Y, Freeling MR, Borm T, Bonnema G, Wu J, Wang X (2016) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. Nature Genetics 48(10):1218–1224.

Chin C, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR and MC Schatz (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* 13, 1050–1054.

Churchill GA, Waterman MS (1992). The Accuracy of DNA Sequences: Estimatic Sequence Quality. Genomics 14 (2): 89–98.

Clark AG (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular biology and evolution, 7(2), 111-122.

Clark AG (2004) The Role of Haplotypes in Candidate Gene Studies. Genetic Epidemiology 27 (4).

Collard BCY, Das A, Virk PS, Mackill DJ (2007) Evaluation of 'quick and dirty' DNA extraction methods for marker-assisted selection in rice (Oryza sativa L.). Plant Breeding 126(1):47-50.

Contreras-Soto, RI, Mora F, de Oliveira MAR, Higashi W, Scapim CA, Schuster I (2017). A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. PloS one, 12(2), e0171105.

D'hoop B (2009) Association mapping in tetraploid potato (PhD thesis). ISBN 978-90-8585-333-6

D'hoop B, Keizer PL, Paulo MJ, Visser RGF, van Eeuwijk FA, van Eck HJ (2014). Identification of agronomically important QTL in tetraploid potato cultivars using a marker–trait association analysis. Theoretical and applied genetics, 127(3), 731-748.

D'Hoop B, Paulo MJ, Kowitwanich K, Sengers M, Visser RGF, van Eck HJ, van Eeuwijk FA (2010) Population structure and linkage disequilibrium unravelled in tetraploid potato. Theoretical and Applied Genetics, 121(6), 1151-1170.

D'hoop B, Paulo MJ, Mank RA, Van Eck HJ, van Eeuwijk FA (2008). Association mapping of quality traits in potato (Solanum tuberosum L.). Euphytica, 161(1-2), 47-60.

D'hoop B, Paulo MJ, Visser RGF, van Eck HJ, van Eeuwijk FA (2011) Phenotypic analyses of multi-environment data for two diverse tetraploid potato collections: comparing an academic panel with an industrial panel. Potato Research, 54(2), 157.

Das S, Vikalo H (2015) SDhaP: Haplotype Assembly for Diploids and Polyploids via Semi-Definite Programming. BMC Genomics (16).

De Koeyer D, Douglass K, Murphy A, Whitney S, Nolan L, Song Y, De Jong WS (2010) Application of high-resolution DNA melting for genotyping and variant scanning of diploid and autotetraploid potato. Molecular Breeding, 25(1), 67.

Delaneau O, Coulonges C, Zagury JF (2008). Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC bioinformatics, 9(1), 540.

Delaneau O, Marchini J, McVean GA, Donnelly P, Lunter G, Marchini JL, ..., Rimmer A (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nature communications, 5, 3934.

Di Guardo M, Bink MC, Guerra W, Letschka T, Lozano L, Busatto N, Poles L, Tadiello LB, Visser RGF, van de Weg E (2017). Deciphering the genetic control of fruit texture in apple by multiple family-based analysis and genome-wide association. Journal of experimental botany, 68(7), 1451-1466.

Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet. Epidemiol. , 2003, vol. 25.

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular biology and evolution, 12(5), 921-927.

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Douches DS (2012). Integration of two diploid potato linkage maps with the potato genome sequence. PloS one, 7(4), e36347.

Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. A*nnual review of plant biology, 64, 267-291.*

Fisher, RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb., 52, 399-433

Flipse E, Keetels CJAM, Jacobsen E, Visser RGF (1996) The dosage effect of the wildtype GBSS allele is linear for GBSS activity but not for amylose content: absence of amylose has a distinct influence on the physico-chemical properties of starch. *Theoretical and applied genetics*, *92*(1), 121-127.

Francoz, Edith et al. 2014. "Roles of Cell Wall Peroxidases in Plant Development." Phytochemistry. http://www.ncbi.nlm.nih.gov/pubmed/25109234 (October 17, 2014).

Frary A, Nesbitt AC, Frary A, Grandillo S, van der Knaap E, Con B, Liu J, Meller J, Elber R, Alpert KB and Tanksley SD (2000) fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size. *Science, Vol. 289, Issue 5476, pp. 85-88*

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2003) The Structure of Haplotype Blocks in the Human Genome. *Science, 296 (5576), 2225-2229*

Gan, X, Stegle, O, Behr, J, Steffen, JG, Drewe, P, Hildebrand, KL, Lyngsoe, R, Schultheiss, SJ, Osborne, EJ, Sreedharan, VT, Kahles, A, Bohnert, R, Jean, G, Derwent, P, Kersey, P, Belfield, EJ, Harberd, NP, Kemen, E, Toomajian, C, Kover, PX, Clark, RM, Rätsch, G, and R Mott. (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477, 419–423

Gao X, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic Epidemiology 32(4): 361-369. 10.1002/gepi.20310.

Garg S, Martin M, Marschall T (2016) Read-based phasing of related individuals. Bioinformatics, 32(12), i234–i242.

Garrison E, Marth G (2012). Haplotype-Based Variant Detection from Short-Read Sequencing. arXiv Preprint arXiv:1207.3907.

Gazzani S, Gendall AR, Lister C, Dean C (2003) Analysis of the molecular basis of flowering time variation in Arabidopsis accessions. *Plant physiology*, *132*(2), 1107-1114.

Gegas VC, Nazari A, Griffiths S, Simmonds J, Fish L, Orford S, ... Snape JW (2010) A genetic framework for grain size and shape variation in wheat. *The Plant Cell*, *22*(4), 1046-1056.

Goddard, ME Hayes BJ (2007) Genomic selection. *J. Anim. Breed. Genet. 124 323–330*

Hackett CA, McLean K, Bryan GJ (2013) Linkage Analysis and QTL Mapping Using SNP Dosage Data in a Tetraploid Potato Mapping Population. PLoS ONE 8 (5).

Hamblin MT, Jannink JL (2011) Factors affecting the power of haplotype markers in association studies. The Plant Genome, 4(2), 145-153.

Hamilton JP, Hansey CN, Whitty, BR, Stoffel K, Massa AN, Van Deynze A, De Jong WS, Douches DS, Buell CR (2011) Single nucleotide polymorphism discovery in elite north american potato germplasm. BMC Genomics 12, 302.

Hardigan MA, Crisovan E, Hamiltion JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B, Yang X, Zeng Z, Douches, DS, Jiang J, Veilleux RE, Buell CR (2016) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. The Plant Cell, 28(2), 388-405.

Hibrand Saint-Oyant L, Ruttink R, Hamama L, Kirov I, Lakhwani D, Zhou NN, Bourke P, Daccord N, Leus L, Schulz D, Van de Geest H, Hesselink T, Van Laere K, Balzergue S, Thouroude T, Chastellier A, Jeauffre J, Voisine L, Gaillard S, Borm T, Arens P, Voorrips R, Maliepaard C, Neu E, Linde M, Le Paslier MC, Berard A, Bounon R, Clotault J, Choisne N, Quesneville H, Kawamura K, Aubourg S, Sakr S, Smulders R, Schijlen E, Bucher E, Debener T, de Riek J, Foucher F (2018) A high-quality sequence of Rosa chinensis to elucidate genome structure and ornamental traits. *Nature Plants (2018)*

Hill, WG, Goddard, ME, and PM Visscher (2008) Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLOSgenetics 4:2*

Hirsch CN, Hirsch CD, Felcher K, Coombs J, Zarka D, Van Deynze A, Douches, DS (2013) Retrospective view of North American potato (Solanum tuberosum L.) breeding in the 20th and 21st centuries. G3: Genes, Genomes, Genetics, g3-113.

Hong H, Xu L, Liu J, Jones WD, Su Z, Ning B (2012) Technical Reproducibility of Genotyping SNP Arrays Used in Genome-Wide Association Studies. PLoS ONE 7(9)

Huang W, Li L, Myers JR, Marth, GT (2011) ART: a next-generation sequencing read simulator. Bioinformatics, 28(4), 593-594.

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annual review of plant biology, 65, 531-551.

Hulse-Kemp AM, Maheshwari S Stoffel K, Hill TA, Jaffe D, Willias SR, Weisenfeld N, Ramakrishnan S, Kumar V, Shah P, Schatz MC, Church DM, Van Deynze A (2018)

Reference quality assembly of the 3.5-Gb genome of Capsicum annuum from a single linked-read library. *Horticulture Research 5:4.*

Ingvarsson PK, Street NR (2011). Association genetics of complex traits in plants. New Phytol 189:909–922

Jiao WB, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. *Current opinion in plant biology, 36, 64-70.*

Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RCJ, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SCL, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics volume 29, pages 233–237.*

Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ (2010) Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics, 11:724

Jong H, Burns VJ (1993) Inheritance of Tuber Shape in Cultivated Diploid Potatoes. American Potato Journal 70(3): 267–84.

Jongedijk E, van der Wolk JMASA, Suurs LCJM (1990) Analysis of glutamate oxaloacetate transaminase (GOT) isozyme variants in diploid tuberous Solanum; inheritance and linkage relationships to ds1 (desynapsis), y (tuber flesh colour), cr (crumpled) and yc (yellow cotyledon). Euphytica 45(2):155-167.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008). Efficient control of population structure in model organism association mapping. Genetics, 178(3), 1709-1723.

Kim JH, Waterman MS, Li LM (2007) Diploid Genome Reconstruction of Ciona Intestinalis and Comparative Analysis with Ciona Savignyi. Genome Research 17 (7): 1101–10.

Kloosterman B, Abelenda JA, del Mar Carretero Gomez M, Oortwijn M, de Boer JM, Kowitwanich K, Horvath BM, van Eck HJ, Smaczniak C, Prat S, Visser RGF, Bachem

CWB (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature, 495(7440), 246.

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant methods 9:1.

Krumsiek J, Arnold R, Rattei T (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23(8):1026–1028.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis, G, Durbin R (2009) The Sequence Alignment/Map Format and SAMtools. Bioinformatics 25 (16): 2078–79.

Li ML, Kim JH, Waterman MS (2004) Haplotype Reconstruction from SNP Alignment. Journal of Computational Biology 11 (2–3): 507–18.

Li X, van Eck HJ, van der Voort R, Huigen D-J, Stam P, Jacobsen E (1998) Autotetraploids and genetic mapping using common AFLP markers: the R2 allele conferring resistance to Phytophthora infestans mapped on potato chromosome 4. Theoretical and Applied genetics 96:;1121-1128.

Li X-Q, De Jong H, De Jong DM, De Jong WS (2005) Inheritance and Genetic Mapping of Tuber Eye Depth in Cultivated Diploid Potatoes. Theor Applied Genet 110(6): 1068–1073.

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. Proc Natl Acad Sci USA. 107 (49)

Lindqvist-Kreuze H, Khan A, Salas E, Meiyalaghan S, Thomson S, Gomez, R, Bonierbale M (2015) Tuber shape and eye depth variation in a diploid family of Andean potatoes. *BMC genetics*, *16*(1), 57.

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A Mixed-Model Approach to Association Mapping Using Pedigree Information with an Illustration of Resistance to Phytophthora Infestans in Potato. Genetics 175(2): 879–89.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 20(9), 1297-1303.

Meuwissen TH, Goddard ME (2000) Fine Mapping of Quantitative Trait Loci Using Linkage Disequilibria with Closely Linked Marker Loci. Genetics 155 (1): 421–30.

Meuwissen, T (2007) Genomic selection: marker assisted selection on a genome wide scale. *Journal of animal Breeding and genetics*, *124*(6), 321-322.

Minio A, Lin J, Gaut BS, Cantu D (2017) How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Frontiers in plant science*, *8*, 826.

Monforte, AJ, Aurora D, Caño-Delgado A, van der Knaap E (2014) The Genetic Basis of Fruit Morphology in Horticultural Crops: Lessons from Tomato and Melon. Journal of experimental botany 65(16): 4625–37.

Morrell PL, Buckler, ES, Ross-Ibarra J (2011) Crop Genomics: Advances and Applications Nature Reviews Genetics 13 (2).

Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genetic epidemiology, 23(3), 221-233.

Mosquera T, Alvarez MF, Jiménez-Gómez JM, Muktar ME, Paulo MJ, Steinemann S, Li J (2016) Targeted and Untargeted Approaches Unravel Novel Candidate Genes and Diagnostic SNPs for Quantitative Resistance of the Potato (Solanum Tuberosum L.) to Phytophthora Infestans Causing the Late Blight Disease. PLoS ONE 11 (6): 1–36.

Motazedi E, de Ridder D, Finkers R, Baldwin S, Thomson S, Monaghan K, Maliepaard C (2018) TriPoly: haplotype estimation for polyploids using sequencing data of related individuals, Bioinformatics, bty442

Motazedi E, Finkers R, Maliepaard C, de Ridder D (2017) Exploiting next-Generation Sequencing to Solve the Haplotyping Puzzle in Polyploids: A Simulation Study. Briefings in Bioinformatics 1;19(3):387-403.

N'Diaye A, Haile JK, Cory AT, Clarke FR, Clarke JM, Knox RE, Pozniak CJ (2017) Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. PloS one, 12(1), e0170941.

Neigenfind J, Gyetvai G, Basekow R, Diehl S, Achenbach U, Gebhardt C, Selbig J, Kersten B (2008) Haplotype Inference from Unphased SNP Data in Heterozygous Polyploids Based on SAT BMC Genomics 9 (1): 356

Nelson RM, Pettersson ME, Carlborg Ö (2013) A century after Fisher: time for a new paradigm in quantitative genetics. *Trends in Genetics, 29:12.*

Orr, HA (2005) The genetic theory of adaptation: a brief history. *Nature Reviews Genetics, 6(2), 119.*

Passardi F, Cosio C, Penel C, Dunand C (2005) Peroxidases Have More Functions than a Swiss Army Knife. Plant Cell Reports 24: 255–65.

Passardi F, Tognolli M, De Meyer M, Penel C, Dunand C (2006) Two cell wall associated peroxidases from Arabidopsis influence root elongation. *Planta*, *223*(5), 965-974.

Patterson M, Marschall T, Pisanti N, van Iersel, L (2014) WHatsHAp: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 22 (0): 1–12.

Pedreira J, Herrera MT, Zarra I, Revilla G (2011) The Overexpression of AtPrx37, an Apoplastic Peroxidase, Reduces Growth in Arabidopsis. Physiologia plantarum 141(2): 177–87.

Pham GM, Newton L, Wiegert-Rininger K, Vaillancourt B, Douches DS, CR Buell (2017) Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. The Plant Journal (2017) 92, 624–637

Potato Genome Sequencing Consortium (2011) Genome Sequence and Analysis of the Tuber Crop Potato. Nature 475(7355): 189–95.

Prashar A, Hornyik C, Young V, McLean K, Sharma SK, Dale MFB and Bryan GJ (2014) Construction of a Dense SNP Map of a Highly Heterozygous Diploid Potato Population and QTL Analysis of Tuber Shape and Eye Depth. Theoretical and applied genetics. 127 (10).

Quinlan AR, Hall IM (2010) BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. Bioinformatics 26 (6): 841–42

Rosyara UR, De Jong WS, Douches DS, Endelman JB (2016) Software for genome-wide association studies in autopolyploids and its application to potato. The plant genome, 9(2).

Salamov AA, Solovyev VV (2000) Ab Initio Gene Finding in Drosophila Genomic DNA. Genome Res. 10(4):516-22

Schaid DJ (2004) Evaluating associations of haplotypes with traits. Genetic epidemiology, 27(4), 348-364.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics, 78(4), 629-644.

Schmidt MHW, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Maß J, Pfaff C, Schurr U, Chetelat R, Maumus F, Aury JM, Koren S, Fernie AR, Zamir D, Bolger AM, Usadel B (2017) De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *The Plant Cell, Vol. 29: 2336–2348.*

Schönhals EM, Ortega F, Barandalla L, Aragones A, De Galarreta JR, Liao J, Gebhardt C (2016) Identification and reproducibility of diagnostic DNA markers for tuber starch and yield optimization in a novel association mapping population of potato (Solanum tuberosum L.). Theoretical and Applied Genetics, 129(4), 767-785.

Schreiber L, Nader-Nieto AC, Schönhals EM, Walkemeier B, Gebhardt C (2014) SNPs in Genes Functional in Starch-Sugar Interconversion Associate with Natural Variation of Tuber Starch and Sugar Content of Potato (Solanum Tuberosum L.). G3 4 (10): 1797–1811.

Schwartz R (2010) Theory and Algorithms for the Haplotype Assembly Problem. Communications in Information and Systems 10 (1): 23–38.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics, 44(7), 825.

Sharma SJ, MacKenzie K, McLean K, Dale,F, Daniels S, Bryan GJ (2018) Linkage Disequilibrium and Evaluation 1 of Genome-Wide Association Mapping Models in Tetraploid Potato (In press).

Shen J, Li Z, Chen J, Song Z, Zhou Z, Shi Y (2016) SHEsisPlus, a Toolset for Genetic Studies on Polyploid Species. Scientific Reports 6.

Simko I (2004) One potato, two potato: haplotype association mapping in autotetraploids. Trends in plant science, 9(9), 441-448.

Sirks MJ (1929) Multiple allelomorphs versus multiple factors. *Proc. of the Int. Congr. of PI. Sei., Ithaca, New York 1:803-814.*

Slater AT, Cogan NO, Hayes BJ, Schultz L, Dale, MFB, Bryan GJ, Forster, JW (2014) Improving breeding efficiency in potato using molecular and quantitative genetics. Theoretical and applied genetics, 127(11), 2279-2292.

Slater AT, Cogan NOI, Forster JW, Hayes BJ, Daetwyler HJ (2016) Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. the plant genome. *The plant genome 9 (3).*

Slater AT, Wilson GM, Cogan NOI, Forster JW, Hayes BJ (2014) Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. *Theor Appl Genet 127:809–820*

Snyder MW, Kitzman AA, Shendure J (2015) Haplotype-resolved genome sequencing: experimental methods and applications. Nature Reviews Genetics, 16(6), 344-358.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 7(Suppl 1): 10.1-10.12.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. The American Journal of Human Genetics, 73(5), 1162-1169.

Stich B, van Inghelandt D (2018) Prospects and Potential Uses of Genomic Prediction of Key Performance Traits in Tetraploid Potato. *Frontiers in plant science, 9, 159.*

Stram, DO Seshan VE (2012) Multi-SNP haplotype analysis methods for association analysis. In Statistical Human Genetics (pp. 423-452).

Su SY, White J, Balding DJ, Coin LJM (2008) Inference of Haplotypic Phase and Missing Genotypes in Polyploid Organisms and Variable Copy Number Genomic Regions. BMC Bioinformatics 9 (1): 513

Su SY, Asher JE, Jarvelin MR, Froguel P, Blakemore AI, Balding, DJ, Coin, LJ (2010) Inferring combined CNV/SNP haplotypes from genotype data. Bioinformatics, 26(11), 1437-1445.

Tang H (2017). Disentangling a polyploid genome. Nature plants, 3

Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JM (2006) QualitySNP: A Pipeline for Detecting Single Nucleotide Polymorphisms and Insertions/deletions in EST Data from Diploid and Polyploid Species. BMC Bioinformatics 7: 438

The 100 Tomato Genome Sequencing Consortium (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *The Plant Journal 80, 136–148.*

The Tomato Genome Consortium (2009) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature (485), 635–641.*

Uitdewilligen JGAML, Wolters AMA, d'Hoop DB, Borm TJA, Visser RGF, van Eck HJ (2013) A next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato. PloS One 8.

Uitdewilligen JGAML (2012) Discovery and genotyping of existing and induced DNA sequence variation in potato. ISBN: 978-94-6173-233-0

Uitdewilligen, JGAML, Wolters, AMA, van Eck, HJ, Visser RGF (2013) Sequence Characterization of StGWD Haplotypes and the Genetics of Starch Phosphate Content in Tetraploid Potato. In: Discovery and Genotyping of Existing and Induced DNA Sequence Variation in Potato. ISBN: 978-94-6173-233-0

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA (2007) Primer3Plus, an Enhanced Web Interface to Primer3. Nucleic acids research 35: W71–74.

Van Berloo R, Hutten RCB, Van Eck HJ, Visser RGF (2007). An online potato pedigree database resource. Potato research, 50(1), 45-57.

Van de Wal MHBK, Jacobsen E, Visser RGF (2001) Multiple allelism as a control mechanism in metabolic pathways: GBSSI allelic composition affects the activity of granule-bound starch synthase I and starch composition in potato. *Molecular Genetics and Genomics*, *265(6), 1011-1021.*

van Eck HJ (1995) Localisation of morphological traits on the genetic map of potato using RFLP and isozyme markers. *PhD thesis*

Van Eck HJ, Jacobs JM, Stam P, Ton J, Stiekema WJ, Jacobsen E (1994).Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics, 137(1), 303-309.

van Eck HJ, Vos PG, Valkonen JP, Uitdewilligen J, LensingH, de Vetten N, Visser, RGF (2017) Graphical genotyping as a method to map Ny (o, n) sto and Gpa5 using a reference panel of tetraploid potato cultivars. Theoretical and Applied Genetics, 130(3), 515-528.

Van Eck HJ, Vos PG, Valkonen JPT, Uitdewilligen JGAML, Lensing H, de Vetten N, Visser RGF (2017) Graphical genotyping as a method to map $Ny_{(o,n)sto}$ and Gpa5 using a reference panel of tetraploid potato cultivars. Theor Appl Genet. 130(3): 515–528.

van Eck, HJ (2007) Genetics of Morphological and Tuber Traits. In Potato Biology and Biotechnology, ed. Taylor MA, Ross HA, Vreugdenhil D, Bradshaw, Gebhardt C, Govers F, Mackerron DKL.

Van Geest G, Bourke PM, Voorrips RE, Marasek-Ciolakowska A, Liao Y, Post A., van Meeteren U, Visser RGF, Maliepaard C, Arens P (2017) An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. Theoretical and Applied Genetics, 130(12), 2527-2541.

van Geest G, Post A, Arens P, Visser RGF, van Meeteren U (2017) Breeding for postharvest performance in chrysanthemum by selection against storage-induced degreening of disk florets. Postharvest Biology and Technology, 124, 45-53.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates. Genome research 19(2): 327–35.

Voorrips RE, Maliepaard C (2008) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC Bioinformatics 13:248.

Voorrips, RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC bioinformatics, 12(1), 172.

Vos PG (2016) Development and application of a 20K SNP array in potato (PhD thesis) IBSN: 978-94-6257-956-9.

Vos PG, Paulo MJ, Bourke PM, Maliepaard CA, Visser RGF, van Eeuwijk FA, van Eck HJ (2016) GWAS in tetraploid potato: Identification and validation of SNP markers associated with glycoalkaloid content. In: Development and application of a 20K SNP array in potato. IBSN: 978-94-6257-956-9.

Vos PG, Paulo MJ, Voorrips RE, Visser RGF, van Eck HJ, van Eeuwijk FA (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theoretical and Applied Genetics, 130(1), 123-135.

Vos PG, Uitdewilligen JG, Voorrips RE, Visser RGF, van Eck HJ (2015) Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. Theoretical and Applied Genetics, 128(12), 2387-2401.

Vukosavljev M, Arens P, Voorrips RE, van 't Westende WP, Esselink GD, Bourke PM, Cox P, Van de Weg WE, Visser RGF, Maliepaard C, Smulders MJM (2016) High-density

SNP-based genetic maps for the parents of an outcrossed and a selfed tetraploid garden rose cross, inferred from admixed progeny using the 68k rose SNP array. Horticulture research (3):16052.

Wall JD,Pritchard JK (2003) Assessing the Performance of the Haplotype Block Model of Linkage Disequilibrium. Am J Hum Genet. 73(3): 502–515.

Witek K, Jupe F, Witek, AI, Baker D, Clark MD, Jones, JD (2016). Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. Nature biotechnology, 34(6), 656.

Wolters AMA, Uitdewilligen JGMAL, Kloosterman BA, Hutten RCB, Visser RGF, van Eck HJ (2010) Identification of Alleles of Carotenoid Pathway Genes Important for Zeaxanthin Accumulation in Potato Tubers. Plant Molecular Biology 73 (6): 659–71

Xie M, Wu Q, Wang J, Jiang T (2016) H-PoP and H-PoPG: Heuristic Partitioning Algorithms for Single Individual Haplotyping of Polyploids. Bioinformatics 32 (24)

Yang Q, Li Z, Li W, Ku L, Wang C, Ye J, Li J (2013) CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the post domestication spread of maize. Proceedings of the National Academy of Sciences, 110(42)

Yu LX, Zheng P , Bhamidimarri S, Xiang-Ping L, Main D (2017) The Impact of Genotyping-by-Sequencing Pipelines on SNP Discovery and Identification of Markers Associated with Verticillium Wilt Resistance in Autotetraploid Alfalfa (Medicago Sativa L.) Frontiers in Plant Science 8: 89.

Zaitlen N, Kraft P (2012) Heritability in the Genome-Wide Association Era. Human Genetics 131(10): 1655–64.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Human heredity, 53(2), 79-91.

Zheng C, Voorrips RE, Jansen J, Hackett CA, Ho J, Bink MCAM (2016) Probabilistic Multilocus Haplotype Reconstruction in Outcrossing Tetraploids. Genetics 203 (1): 119–31.

Summary

The identification of haplotypes in tetraploid potato allows to improve genetic studies and facilitate marker-assisted selection. For many years, only bi-allelic molecular markers were used for application in genetic studies and they undoubtedly improved our understanding of the inheritance of important agronomical traits. However, these undertakings are complicated by the lack of knowledge about linkage between these SNPs and thus their underlying haplotype structure. The inability of geneticists to achieve haplotype reconstruction was mainly due to complications of the higher ploidy level of cultivated potato (2x = 4x = 48), as a single potato variety contains four copies of each chromosome (tetraploid). In this thesis, methods are described that allow haplotype reconstruction in tetraploid potato, either from sequencing data of a single variety or by use of SNP information over multiple varieties. We employed these methods on genotypic data of potato varieties and used the reconstructed haplotypes to detect which alleles influence traits such as plant maturity, tuber shape and flesh color.

The starting point of this thesis was a genetic study of the inheritance of potato tuber shape and eye depth. In **Chapter 2** we identified a strong marker-trait association for tuber shape on potato chromosome *10* (*Ro* locus), that co-localises with a major effect QTL for eye depth. Subsequent fine mapping in a diploid full-sib potato population (C × E) refined the associated region of 3.1 Mb to a small region of 280 Kb. In this region, a repeat cluster of peroxidase genes is located.

In **Chapter 3** we started with the development of methods for haplotype reconstruction. We introduced a novel method to use short-read DNA sequencing data to reconstruct haplotypes. A previous study genotyped ~800 potato genes in 83 tetraploid varieties using Illumina short reads. This information was used as input for our haplotype reconstruction pipeline and allowed us to generate haplotype blocks of 413 bp average in tetraploid potato, and estimate the haplotype diversity in potato. In addition, we performed a simulation study, which showed that our approach had superior accuracy compared to competing approaches.

A disadvantage of haplotype reconstruction with sequencing data is that only short-range haplotypes can be reconstructed. To facilitate the construction of long-range

haplotypes, we developed in **Chapter 4** a method that allows estimating haplotypes on basis of genetic information over multiple samples. This was achieved by first reconstructing linkage phase between SNP pairs, followed by the joining of these linkage phases into full-length haplotypes. We validated this method by use of pre-existing haplotypes of the *StGWD1* gene. This validation study indicated that haplotype reconstruction is highly accurate. In addition, we employed our method on genotypic data of potato. The results show that the haplotype diversity in potato is extensive, but that a few common haplotypes are responsible for the majority of allelic variation.

In **Chapter 5** we subsequently used these haplotypes to explore the application of haplotypes in a haplotype-based GWAS. Conventionally, GWAS is only performed with bi-allelic SNP markers, but knowledge of haplotype-specificity is required to interpret the resulting marker-trait associations. Here we performed haplotype-based GWAS and compared this to the results of single marker GWAS. We linked specific alleles to potato traits such as plant maturity, tuber shape, flesh color and potato tuber uniformity.

In **Chapter 6** we report the development of Poly-Imputer. This tool allows to perform haplotype imputation and is based on the intuition that if the most or all segregating alleles are known it becomes trivial to assign four of these haplotypes to any individual. As input, we used a library of reference haplotypes and dosage calls of each variety. Application of this tool allowed to perform phasing of SNPs in progeny of a full-sib population, but more importantly also refine and improve haplotype solutions that are reconstructed with sequencing data and haplotypes based on dosage data.

Chapter 7 involves the determination of haplotype diversity at the *StCDF1* gene, a key regulator of the tuberization response in potato. In this study, we performed haplotype assembly for the 2nd exon of this gene, followed by manual assignment of haplotypes by use of sequencing reads and genetic relations. In this study, we could demonstrate a significant phenotypic effect of only one *StCDF1* allele.

In the final chapter, we discuss the findings of the previous six chapters. In conclusion, this thesis provides a significant step for routine investigation of haplotype diversity in tetraploid potato. Hopefully, the methods and tools provided in this thesis will facilitate the use of haplotypes in marker-assisted selection and increase our understanding of allele-phenotype interactions in potato.

Acknowledgements

After four years of hard work, this booklet is finally completed. The results in this thesis could not have been achieved without the help of others, for which this section of the PhD thesis gives ample opportunity to express my feelings of gratitude.

First of all, I wanted to thank my daily supervisor, **Herman van Eck**. You are an excellent scientist from which I learned a lot during the past four years. We first met during a MSc-thesis project, which subsequently evolved in a full PhD project. Not everything that was initially planned delivered the results we hoped for, but along the road we learned a lot. You gave me a lot of freedom to explore my own ideas for research, which is something that I liked a lot. You are a very dedicated scientist and I admire your sharp eye for detail. During the years of working together I found out that this is sometimes lacking in my writing. However, I am happy to have learned from the best. I sincerely appreciate the discussions about all things that are above, and beside science. On behalf of my children you need a word of thank for the 'glijbaan' that has become a central and necessary element of our garden.

To **Richard Visser**, my promoter. Thank you for letting me pursue a PhD at plant breeding, and critical assessment analysis of my chapters, but also to inspire me to finish my PhD within the four years. You are very objective, but also very pragmatic in your approach, for which I want to thank you.

A very special thanks to **Michiel Klaassen**. We shared many things over the years during all the coffee and lunch breaks. Thanks for all the support! Also, thanks for being my paranimph. We went together on a trip to PAG in San Diego, which was a lot of fun. You always have a very positive vibe around you. I am very confident you will finish your PhD soon and wish you all the best.

Likewise I have to thank **Geert van Geest** for being my paranimph. We worked alongside for several years, while doing research in very different crops. You are the fastest implementer I know of, and I look forwards to further endeavours on the polyploid road.

I would also like to thank all others who have been involved in my project: **Theo Borm**, thanks for the many discussions about haplotyping and sequencing and for letting me work on the plant breeding HPC. **Richard Finkers** for all suggestions and help. **Christian Bachem**, you were very instrumental in the many discussions that we had about the *StCDF1* gene, which greatly helped my understanding of molecular biology.

A special thanks to the people of the Biodiversity and quantitative genetics group. First of all, **René Smulders** for suggestions and feedback. Likewise, I would like to thank **Chris Maliepaard, Roeland Voorrips, Ronald Hutten** and **Eric van der Weg**. A special thanks to **Paul Arens** for letting me set up a capture-seq experiment in three polyploid corps. It is unfortunate that this could not be part of this thesis, but hopefully it will in the future lead to (a) interesting publication(s).

Besides people at plant breeding, also the people involved in my project from the potato companies are greatly appreciated for their contribution: **Nick de Vetten, Emmet Dalton, Guus Heselmans, Jan-David Driesprong, Johan Hopman, Jan de Boer** and **Jan de Haas**.

A special thanks to my regular lunch (and coffee) colleagues: **Peter Bourke, Thijs van Dijk, Geert van Geest, Michiel Klaassen** and **Peter Vos.** I would also like to thank my office mates, both former and current: **Ashikin, Ernest, Elisabeth, Jeroen, Kaile, Peter Dihn** and **Xinfang.** ●ther people that need a word of thank are: **Arwa, Cynara, Ehsan, Gurnoor, Giorgio, Jarst, Johan B, Mas, Pauline, Sri** and **Yanlin.** Also my two former MSc students **Jean Custers** and **Yiyuan Ding** are acknowledged for their help. I realize now that I am likely to forget people, so beforehand my sincere apologies for that.

Next, I am going to switch to Dutch: De meeste woorden van dank zouden moeten uitgaan naar mijn familie en vrienden. Jullie waren altijd erg betrokken bij mijn onderzoek. Dat heb ik altijd gewaardeerd. Natuurlijk was er zonder de hulp van Leanne niks van dit proefschrift terecht gekomen. Dankjewel voor al het geduld!

About the author

Johan (Johannis Hendrikus) Willemsen was born on 21th July 1991 in Wageningen, the Netherlands. In 2009 he finished his secondary education (VWO). His interest in plants was raised while working in a tree nursery, and in 2009 he started his study plant sciences at the Wageningen university. During his MSc, he performed a minor thesis at the laboratory of Genetics under supervision of Prof. Dr. Ir. M.G.M. Aarts, exploring the use of the haploid moss *Physcomitrella patens* for use in research into zinc deficiency. In that period, he developed also an interest in science, and more particular in (plant) genetics and bioinformatics. In spring 2014, his major thesis was completed in the department of Plant Breeding under supervision of Dr. Ir. H.J. van Eck, with as topic fine mapping of a major effect QTL for potato tuber shape. In November 2014 he subsequently graduated with a master degree in plant sciences with the specialization Plant breeding and Genetic resources.

Meanwhile, during his MSc he wrote a grant application for the NWO-ALW about the discovery of allelic diversity in potato. After a successful defence of the proposal, he started as a PhD student at the department of Plant Breeding under the supervision of Dr. ir. H.J. van Eck and Prof. Dr. R. G. F. Visser. This research focussed on the development of tools to investigate the allelic composition of tetraploid potato varieties. These results are presented in this PhD thesis. After four years working with a tetraploid crop he decided to switch efforts to the allooctaploid strawberry and joined Fresh Forward, Eck en Wiel, the Netherlands, as a bioinformatician.

Education statement

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: **Johan H. Willemsen**
Date: **21 November 2018**
Group: **Laboratory of Plant Breeding**
University: **Wageningen University & Research**

	date
1) Start-up phase	
► **First presentation of your project**	
Using NGS data to reconstruct haplotypes in potato: Progress and challenges	29 Jun 2015
► **Writing or rewriting a project proposal**	
The identification of natural variation related to organ development: Plant maturity and tuber shape	01 Nov 2016
► **Writing a review or book chapter**	
► **MSc courses**	
Programming in Python INF-22306	02 okt 2014
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*7.0 credits **

	date
2) Scientific Exposure	
► **EPS PhD student days**	
EPS PhD Student Day 'Get2Gether', Soest, NL	28-29 Jan 2016
EPS PhD Student Day 'Get2Gether', Soest, NL	09-10 Feb 2017
► **EPS theme symposia**	
EPS Theme 4 Symposium 'Genome Biology', Amsterdam, NL	15 Dec 2015
EPS Theme 1 Symposium 'Developmental Biology of Plants', Wageningen, NL	21 Jan 2016
EPS Theme 4 Symposium 'Genome Biology', Wageningen, NL	16 Dec 2016
► **National meetings (e.g. Lunteren days) and other National Platforms**	
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	13-14 Apr 2015
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	11-12 Apr 2016
Consortium meeting: Polyploids, Wageningen, NL	23 Feb 2017
Consortium meeting: Potato, Wageningen, NL	23 May 2017
Consortium meeting: Polyploids, Wageningen, NL	11 Apr 2018
► **Seminars (series), workshops and symposia**	
Symposium: Plant Breeding Research Day 2014, Wageningen, NL	23 Sep 2014
Symposium: From big data to biological solutions, Wageningen, NL	18 Jun 2015
Symposium: WURomics symposium, Wageningen, NL	15 Sep 2016
Symposium: The Future of Genomics, Leiden, NL	31 Oct 2017
Symposium: Access to (plant) data, Wageningen, NL	11 Dec 2017
Symposium: Network Event TKI Horticulture & Propagation Materials, Nieuwegein, NL	03 Apr 2018
Symposium: EPS Polyploidy genetics and breeding, Wageningen, NL	14 Jun 2018
Seminar: Genetic analysis in MAGIC: advantages and challenges (Emma Huang)	25 Jun 2014
Seminar: Seasonal flowering in annual and perennial plants (George Coupland)	19 Jan 2015
Seminar: The evolutionary significance of gene and genome duplications (Yves van de Peer)	03 Feb 2015
Seminar: And yet they oscillate: functional analysis of circadian long non-coding RNAs (Rosanna Henriques)	16 Nov 2015
Seminar: How jasmonates provide the key to harness plant chemistry (Alain Goossens)	08 Dec 2015
Seminar: Genomics-enabled natural products discovery (Douglas Mitchell)	31 Mar 2016
Seminar: Genomic tools to improve disease resistance in spinach (Jim Correll)	03 Jun 2016
Seminar: Salt tolerance in quinoa (Robert van Loo)	03 Jun 2016
Seminar: From QTLS to routine DNA-informed breeding: prospects, advances, and experiences in apple at Washington State University (Cameron Peace)	16 Nov 2016
Seminar: Automated tetraploid genotype calling and its application to pedigree reconstruction in potato (Endelman)	16 Nov 2016
Seminar: Genome-wide association analysis and prediction in tetraploid potato (Endelman)	18 Nov 2016
Seminar: Nanopore Sequencing (Olivier Lucas)	04 Dec 2017
► **Seminar plus**	
► **International symposia and congresses**	
XIV Solanaceae and III Cucurbitaceae Joint Conference, Valencia, Spain	03-06 Sep 2017
XXVI Plant and Animal Genome (PAG), San Diego, USA	12-17 Jan 2018
► **Presentations**	
Poster: Fine mapping of the Ro-locus involved in tuber shape on potato chromosome 10 - Annual meeting 'Experimental Plant Sciences'	11 Apr 2016
Poster: Haplotype reconstruction in potato using next-generation sequencing data - Annual meeting 'Experimental Plant Sciences'	11 Apr 2016
Poster: Haplotype Inference Using SNP Data in Polyploid Potato - XXVI Plant and Animal Genome (PAG)	12 Jan 2018

Poster: Exploiting Short Read Sequencing for the Characterization of Haplotypes in Autotetraploid Potato - XXVI Plant and Animal Genome (PAG)	13 Jan 2018
Talk: Identifying allelic diversity in the autotetraploid potato with sequencing data - WURomics symposium	15 Sep 2016
Talk: Using short read sequencing data to reconstruct haplotypes in potato - WUR B-wise seminars	01 Nov 2016
Talk: Targeted resequencing using bait capture - Polyploids Consortium Meeting	23 Feb 2017
Talk: Allelic diversity in potato - Potato Consortium Meeting	23 May 2017
Talk: Application of targeted resequencing in potato, Chrysanthemum and Alstroemeria - Polyploids Consortium Meeting	11 Apr 2018
► **IAB interview**	
► **Excursions**	
Subtotal Scientific Exposure	*18.4 credits **

3) In-Depth Studies	*date*
► **EPS courses or other PhD courses**	
Linux/Unix scripting course	18 Apr 2018
R for big data science, Wageningen, NL	09-10 May 2016
► **Journal club**	
Participation in literature discussion group at Plant Breeding	2015-2018
► **Individual research training**	
Subtotal In-Depth Studies	*4.5 credits **

4) Personal development	*date*
► **Skill training courses**	
Competence Assesment, Wageningen, NL	04 Nov 2014
Career Perspectives, Wageningen, NL	Sep - Oct 2017
Reviewing a Scientific Paper, Wageningen, NL	15 Mar 2018
Scientific Publishing, Wageningen, NL	05 Apr 2018
► **Organisation of PhD students day, course or conference**	
Organisation of work discussion Plant Breeding - Biodiversity group	2014-2018
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*3.5 credits **

TOTAL NUMBER OF CREDIT POINTS	***33.4 credits ****
Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits. ** A credit represents a normative study load of 28 hours of study.*	

The research described in this thesis was performed at the department of Plant Breeding of Wageningen University & Research. This project was financially supported by the Dutch National Organization for Scientific Research (NWO), under project 831.14.002. The potato breeding companies Averis Seeds B.V., HZPC Holland B.V., KWS POTATO B.V. and Meijer B.V. are also acknowledged for their contribution.

Thesis cover: Several potatoes placed in triangular fashion on a flat surface.

Cover design: Johan Willemsen

Thesis layout by author

Printed by: ProefschriftMaken || www.proefschriftmaken.nl

Financial support from the Department of Plant Breeding (WUR) for printing this thesis is gratefully acknowledged.

Propositions

1. Disentangling the contribution of multiple alleles at a single locus can only be achieved with knowledge of the causative mutations.
   (this thesis)

2. The predictive value of a molecular marker for breeding can only be known when it is clear which alleles are distinguished and which are not.
   (this thesis)

3. Reproducibility of computational research can only be achieved with publication of peer-reviewed computational notebooks.

4. Decarbonisation of the energy system of the Netherlands is only possible if hydrogen is adapted as a fuel source.

5. Effective knowledge valorisation in public-private partnerships can only be achieved by establishment of strong networks between all partners.

6. The societal pre-conception that scientists live in 'ivory towers' will not change with the increased presence of scientists on social media.

7. Compulsory coffee breaks will improve the success of all PhD research.

Propositions belonging to the thesis, entitled:
"**The identification of allelic variation in potato**"
Johan H. Willemsen
Wageningen, 21 November 2018