Development and application of a 20K SNP array in potato

Peter Vos

Thesis Committee

Promotors
Prof. Dr R.G.F. Visser
Professor of Plant Breeding
Wageningen University

Prof. Dr F.A. van Eeuwijk
Professor of Applied Statistics
Wageningen University

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

Other members
Prof. Dr B.J. Zwaan, Wageningen University
Dr J. Endelman, University of Wisconsin-Madison, USA
Dr W.H. Lindhout, Solynta BV, Wageningen
Prof. Dr P.C. Struik, Wageningen University

This research was conducted under the auspices of the graduate school of Experimental Plant Sciences (EPS)

Development and application of a 20K SNP array in potato

Peter Vos

Thesis
submitted in fulfilment of the requirement for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Thursday 17 November 2016
at 11 a.m. in the Aula.

Peter Vos
Development and application of a 20K SNP array in potato
166 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summary in English
DOI: http://dx.doi.org/10.18174/392278
ISBN: 978-94-6257-956-9

Table of contents

Chapter 1

General introduction

Population growth in the 21st century will have great consequences on food security and the environmental impact of food production. The United Nations estimate that in 2050 9.7 billion people are living on this planet (United Nations; Department of Economic and Social Affairs 2015). These 2.3 billion extra compared to the 7.4 billion people living on earth today (2016) will all need food. In addition increasing welfare will put even more pressure on food production in the coming decades. In recent years the world's food production was able to keep up with population growth. Since 2000 the production of the four major staple crops (maize, rice, wheat and potatoes) has grown with almost 37% from 2.10 billion tons to 2.88 billion tons (faostat3.fao.org). Of these four crops maize realized the biggest growth with 72%, while the worldwide potato production only increased 18% between 2000 and 2014. To meet the future need for food, the food production has to keep growing. Nevertheless in the study of Ray et al. (2013) it is stated that the current annual increase of production of four major crops (maize, wheat, corn, soybean) will not meet the need for food in 2050. Possibly potato can play a role to fill this gap, because potato has higher yields per hectare and requires less water per kilo production compared to the other staple crops (Hoekstra and Chapagain 2011). Already in China the government urges the people to grow potatoes because of this. This means that potatoes are grown more often in areas where they were not grown before. Therefore new and well-adapted varieties are needed. Potato breeding can play an important role in developing varieties adapted to the new growing areas. On the other hand worldwide potato breeding has still a major focus on developing varieties for the existing growing areas, which also need better performing varieties. To achieve an increase in potato production for new and existing growing areas is a challenging job where new breeding technologies such as marker-assisted breeding may play an important role.

Genetic gain

In general the increase of crop production can be achieved in two ways, firstly by better agronomical practices and secondly by introducing genetically superior varieties. The latter can be defined as genetic gain, which is often expressed in yield increase per year and frequently illustrated with the corn example as shown in **Fig. 1** (Source: USDA-NASS. 2005). This is a textbook example of how a combination of agronomical practices and genetic gain can improve crop production dramatically. In contrast to the corn example other crops did not make a similar progress as shown in **Fig. 1.** The genetic gain of several crops has been estimated (Higashide and Heuvelink 2009; Rijk et al. 2013; Laidig et al. 2014). For example Higashide and Heuvelink (2009) show that for tomato an average of 0.9% genetic gain for yield is established in the last 30 years. The study of Laidig et al. (2014) compares the yields of varieties released in the last three decades of 12 crops (potato not included) in Germany, showing genetic gain for the majority of these crops. In contrast to Laidig et al. (2014) the study of Rijk

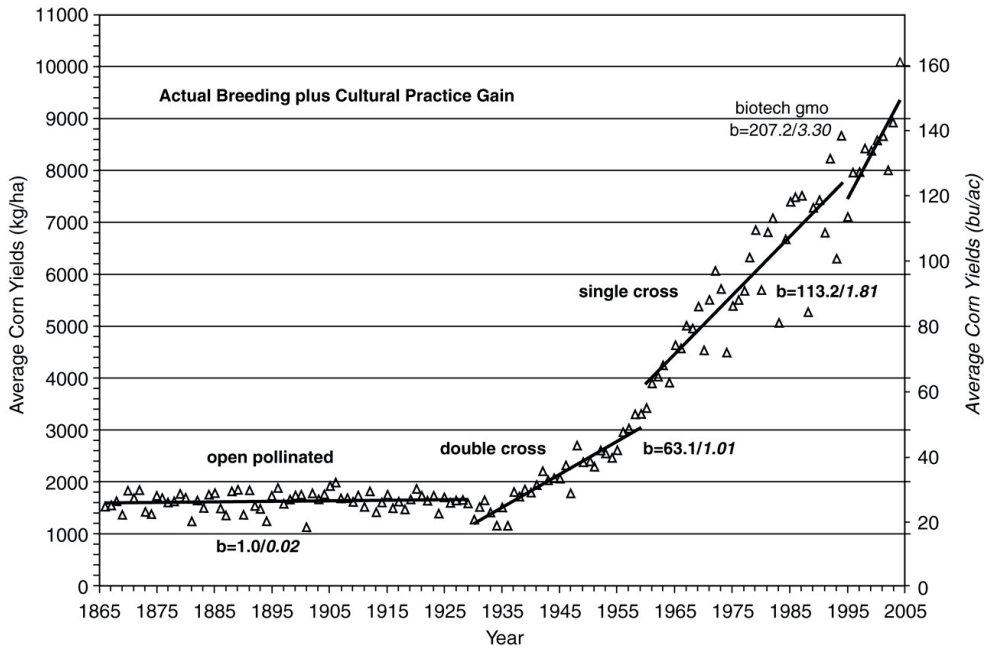

Fig. 1

Increase of corn yields in 140 years. Source: USDA-NASS (2005)

et al. (2013) included potato and performed a similar study for Dutch varieties and showed a limited increase of potato yield as compared to the increase of yield for wheat, barley and sugar beet. Particularly sugar beet displays an exponential increase in crop production over the last decade. The limited genetic gain for yield in potato as reported in Rijk et al. (2013) are supported by data collected within the research of D'hoop (2009) (**Fig. 2a**). In contrast to other crops the genetic gain for yield in potato is limited or even absent. Nevertheless we observed genetic gain in potato for uniformity of yield (uniformity of size and shape). In **Fig. 2b** it is visible that the uniformity of yield (data also collected within the research of D'hoop (2009) is increasing over time, in other words the percentage of marketable yield becomes bigger.

Potato breeding

The limited genetic gain for yield in potato can be explained by several factors specific for potato as a species, but also by the potato market. Firstly the market of potato is much more fragmented compared to other crops such as wheat, maize and sugar beet. For major staple crops such as maize and wheat dry matter or grain yield per hectare are by far the most important traits. In potato many additional traits are of importance, mainly concerning tuber quality. In addition to a higher emphasis on quality traits the different market segments such as the starch industry, French fry industry, crisping

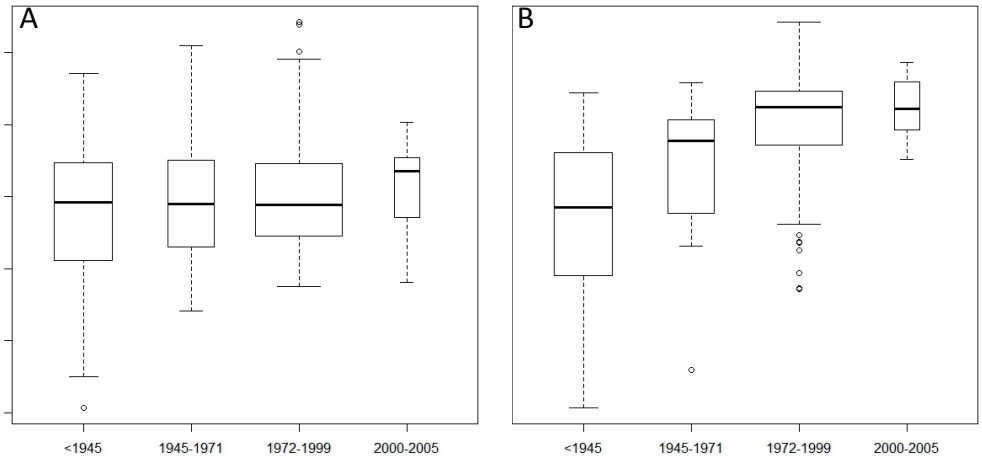

Fig. 2:

Genetic gain in potato. The variety panel described in D'hoop et al. (2008) was divided in four age groups to visualize genetic gain for (A) yield and (B) uniformity of yield.

industry and the fresh consumption require different potato ideotypes. These different market-segments may even require opposite phenotypic trait values for the same trait. A potato breeding program has therefore to consider many more traits. As a consequence, the genetic increase of a single trait like yield lags behind as compared to other crops.

Secondly potato is a highly heterozygous, vegetatively propagated, outcrossing and autotetraploid ($2n = 4x = 48$) species. These four characteristics of potato limit the genetic gain in potato. Firstly the vegetative propagation does not allow rapid multiplication of seed-potatoes; therefore it takes at least 5 years before enough seed tubers are available for reliable (yield) trials. Consequently it takes time to collect sufficient information to decide if a genotype is good enough to use in the next breeding cycle. The prolonged breeding cycle implies fewer meioses over time as compared to annual crops. Typically, some contemporary varieties are only up to 6 or 7 meiotic generations apart from the Rough Purple Chili (Van Berloo et al. 2007), which is considered as one of the founders of the contemporary potato gene pool. Secondly the combination of the heterozygosity, the outcrossing nature and tetrasomic inheritance of potato makes it very difficult to reach fixation of (combinations of) favourable alleles in varieties and advanced breeding clones, which are generally used in a potato breeding program. Such a conventional breeding program of potato consists of multiple F_1 families from crosses between multiple highly heterozygous tetraploid parents. Because favourable alleles are not fixed (homozygous) in parental lines a tremendous segregation for a large number of alleles and a large number of traits is observed, as compared to selfing and/or diploid crops. Only a very small number of offspring will outperform the parents for all traits and

therefore genetic gain is limited. To illustrate the heterozygosity in potato (Uitdewilligen 2012) reported an average of 3.1 different alleles for one gene in one single variety. Taking it all together the time between making a cross and introduction of a new variety takes on average 10-12 years, ranging between 8 and 19 years (Bradshaw 2009).
In general the breeding of new potato varieties relies on several years of phenotypic field selections, performed within the multiple segregating F_1 families until one or two new varieties are released. An alternative approach was proposed by Bradshaw et al. (2003), who propose to discard complete segregating F_1 families before starting selection within F_1 families. However, such an approach is not implemented in Dutch potato breeding practises. More recently different approaches to increase genetic gain in potato have been proposed. For example the use of diploid inbreds to create F_1 hybrids (Lindhout et al. 2011) or EBV (estimated breeding values) based on pedigree information (Slater et al. 2014b) and how MAS (marker assisted selection) can result in genetic gain (Slater et al. 2014a). Marker assisted selection (MAS) is considered as an important tool to increase breeding efficiency and increase genetic gain in general. For almost three decades the scientific literature boasts on the potential of marker assisted breeding, but a broad application in potato stays behind, as compared to other crops such as maize and tomato.

Molecular markers
Marker assisted selection requires molecular markers linked to phenotypic variation and a high throughput system for cost effective data collection in a large breeding program. In this paragraph I describe the shift from low throughput single locus marker systems towards high throughput genotyping assays.
Since the late eighties molecular markers are applied in genetic research in potato. Several techniques and marker systems were developed and applied. Initially single locus marker-systems such as RFLP were implemented resulting in the first RFLP-based potato map (Bonierbale et al. 1988). Later on single sequence repeat (SSR) markers were developed for potato (Milbourne et al. 1998). This genetic resource is widely used, mainly for variety identification for example in Ghislain et al. (2004) but also in association studies (D'hoop et al. 2014). The SSR marker system is a highly informative marker system, because it is multi-allelic. However, it is single PCR-based and therefore not suitable for high-throughput genotyping. In 1995 the AFLP marker system became available (Vos et al. 1995). This marker system allowed researchers to generate dozens of markers per autoradiogram. As a consequence marker density in genetic maps increased, leading to a ultra-dense map of potato containing 10,000 markers (Van Os et al. 2006). A disadvantage of the AFLP method is that markers are randomly distributed over the potato genome. With methylation sensitive restriction enzymes such as PstI these markers can be targeted to gene-rich chromosome-arms, but still a highly over-saturated coverage with markers was obtained for the locus that represented

the pericentromeric heterochromatin (Van Os et al. 2006), which covers almost 50% of the potato genome (Sharma et al. 2013; Bourke et al. 2015). Unfortunately, important markers from an AFLP experiment are not easily converted into single locus assays for practical use in marker-assisted breeding. The underlying polymorphisms in the AFLP system, predominantly single nucleotide polymorphisms (SNP), are the current generation of molecular markers. New high-throughput sequencing technologies allowed identification of vast amounts of SNP-markers (Hamilton et al. 2011; Uitdewilligen et al. 2013). The SNPs from these genotyping studies are publically available and have a fixed location on the reference genome of potato (PGSC 2011). Subsequently these SNP can be used in several types of flexible genotyping platforms, such the Golden Gate system (Anithakumari et al. 2010), the KASP genotyping platform (http://www.kbioscience.co.uk) (Lindhout et al. 2011) and the development of SNP arrays (Felcher et al. 2012).

QTL-Mapping

Molecular markers allowed the development of genetic maps and therewith the identification of genomic regions associated with phenotypic variation. This initially resulted in the identification of loci involved in typical Mendelian traits such as flower colour (van Eck et al. 1993) and tuber skin colour (van Eck et al. 1994). These monogenic traits are easy to phenotype and therefore were very suitable for the primary mapping studies as a proof of concept, however these are typical traits for which MAS is not needed. Mapping studies in the same period in the early nineties resulted in the identification of dominantly inherited resistance genes such as R1 and R3 against *Phytophthora infestans* (El-Kharbotly et al. 1994), H1 against *Globodera rostochiensis* (Gebhardt et al. 1993), followed by the QTLs *Gpa5* and *Gpa6* conferring resistance to *Globodera pallida* (Rouppe van der Voort et al. 2000). These four resistance genes are widely present in the contemporary potato varieties.

To avoid the complex analyses of tetraploid mapping populations the majority of QTL mapping studies are performed in diploid bi-parental mapping population. Diploid mapping populations were either generated from diploid parents, which by prickle pollination and parthenogenetic development of the unfertilized eggs were extracted as primary dihaploids from tetraploid potato varieties (El-Kharbotly et al. 1994; Rouppe van der Voort et al. 2000; Song et al. 2005) or from crosses with diploid parents from the diploid gene pool comprising both cultivated land races and wild species for example in Yencho et al. (1998). QTL mapping in diploid bi-parental population potato was further extended to quantitatively inherited traits such as glycoalkaloid content (Yencho et al. 1998), yield and starch yield (Schäfer-Pregl et al. 1998), cold induced sweetening (Menéndez et al. 2002) and enzymatic discolouration (Werij et al. 2007).

Meanwhile progress was made in the understanding of mapping in tetraploids and consequently mapping studies in tetraploid mapping populations start to be performed.

For example the resistance genes *H3* (Bryan et al. 2002), *R2* (Li et al. 1998) and Ry_{sto} (Brigneti et al. 1997) were all mapped in a tetraploid population. Later on also for quality and agronomic traits tetraploid bi-parental populations were used (Bradshaw et al. 2008; McCord et al. 2011). The majority of studies mentioned above (diploid and tetraploid) have used RFLP or AFLP as genotyping platform. The map-positions of these marker systems are difficult to compare between studies and difficult to compare with the reference genome of potato and are therefore not straightforward to implement in potato breeding.

The study of Anithakumari et al. (2010) firstly uses SNPs (single nucleotide polymorphisms) on important mapping populations such as SH 83-92-488 x RH 89-039-16, used in Van Os et al. (2006) and C x E used in (van Eck et al. 1993; van Eck et al. 1994; Werij et al. 2007; Kloosterman et al. 2010; Wolters et al. 2010). They mined available EST (expressed sequence tags) sequences, to identify SNPs. EST resources were also used by Hamilton et al. (2011), from EST databases from the varieties Bintje, Shepody and Kennebec, in addition they re-sequenced three more varieties (Atlantic, Snowden and Premier Russet) to develop the 8303 SolCAP SNP array (Felcher et al. 2012). The vast amounts of genome wide SNP markers currently available indeed allow high throughput genotyping and therefore should increase the efficiency of the detection of marker trait associations and the application of MAS.

Association Mapping

As a next step in QTL discovery the concept association mapping has been implemented in potato research. In contrast to bi-parental populations, association mapping uses a set of existing varieties or accessions. Association mapping has several advantages as compared to bi-parental populations. Firstly, more genetic variation can be analysed simultaneously, and the mapping resolution is increased because all historical recombination can be exploited. Initial association mapping had a targeted approach on several candidate genes (Li et al. 2008; Li et al. 2010; Baldwin et al. 2011; Urbany et al. 2011; Schreiber et al. 2014). Later the studies of (D'hoop et al. 2008; D'hoop et al. 2014) used an untargeted approach with AFLP markers. The number of markers used in these studies is still limited, therefore a true genome wide association study is the next step in potato research in order to find more QTL regions in one single analysis. Such a true genome wide association study in potato was first published by Uitdewilligen et al. (2013) using a vast amount of 129,156 markers generated by GBS (genotyping by sequencing) on a relative small set of 83 varieties. Larger variety panels were used in GWAS using the SolCAP 8303 array and are only published recently (Mosquera et al. 2016; Rosyara et al. 2016).

Scope of this Thesis

In this thesis I describe the development of the 20K SolSTW SNP array as the most important technical innovation of the last years. With this array a large set of the potato varieties was characterized. Furthermore the array was also used to perform an extensive Genome Wide Association Study (GWAS) using the trait Total glycoalkaloid content as an example.

In **Chapter 2** the development of a 20K SolSTW SNP array is described. Based on a genotyping by sequencing (GBS) dataset of 129,156 markers (Uitdewilligen et al. 2013) a selection was made of 15123 SNPs by removing redundancy and selecting markers without too many flanking SNP. In addition 4463 SNPs were taken from the widely applied SolCAP 8,303 array (Hamilton et al. 2011; Felcher et al. 2012). As a next step a set of 569 potato genotypes was hybridized on this SNP-array. This set of genotypes includes heirloom varieties, important progenitors, varieties from many different countries and several diploids. The inclusion of many old varieties and important progenitors allowed us to identify the effects of breeding efforts of the last 75 years on the composition of the potato gene pool. SNPs could be dated resulting in pre-1945 and post-1945 SNPs, and furthermore selection, founder effects and genetic erosion could be analysed using allele frequency changes over time.

In **Chapter 3** two important factors (population structure and linkage disequilibrium) affecting the outcome of genome wide association studies are analysed. The decay of linkage disequilibrium (LD), determining the resolution of GWAS is described. To understand the patterns of LD in real data several datasets differing in the amount of variation and in the percentage of haplotype specific SNPs were simulated. This resulted in the identification of the 90% percentile and $LD_{1/2,90}$ as being optimal for estimation of LD-decay in potato. Decay of LD was estimated around 2 Mb distance. Population structure was analysed resulting in three major groups, which are mainly the result of breeding toward the processing and starch industry, and a rest group composed of heirloom and modern ware potato.

In **Chapter 4** the potential of GWAS in tetraploid potato was analysed using total glycoalkaloid content (sum of α-solanine and α-chaconine) and the ratio between α-solanine and α-chaconine. The total glycoalkaloid content was highly confounded with population and is illustrative of aspects that may explain the phenomenon of missing heritability. The ratio between α-solanine and α-chaconine was not confounded with population structure and therefore two highly significant QTL, identified near two major candidate genes (*SGT1* & *SGT2*), collectively explained 60% of the phenotypic variance (H^2 =56%).

In **Chapter 5** the panel of 83 varieties and 129,156 SNP markers was used to demonstrate the method of graphical genotyping in a tetraploid variety panel. Graphical genotyping is originally a method to visualize linkage and recombination events from diploid segregating population, but with some modifications it works also very good to visualize introgression segment in a panel of tetraploid varieties. Using this method we identified introgression segments from *Solanum stoloniferum* containing a resistance against Potato Virus Y and a large introgression from *Solanum vernei* containing the *Gpa5* resistance gene.

In the final chapter I discuss the results described in chapters 2,3,4 and 5, and how these results can be used in future research and in marker assisted breeding to accelerate genetic gain in potato.

Chapter 2

Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight in the breeding history

Peter G. Vos[1, 2], Jan G.A.M.L. Uitdewilligen[1], Roeland E. Voorrips[1], Richard G.F. Visser[1, 2], Herman J. van Eck[1, 2]

[1] Plant Breeding, Wageningen University & Research , P.O. Box 386, 6700 AJ Wageningen, The Netherlands
[2] Centre for BioSystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands

Supplementary files can be downloaded from DOI: 10.1007/s00122-015-2593-y
Published in Theoretical and Applied Genetics December 2015, Volume 128, Issue 12, pp 2387–2401

Abstract

A non-redundant subset of 15,138 previously identified SNPs and 4454 SNPs originating from the SolCAP project were combined into a 20k Infinium SNP array for genotyping a total of 569 potato genotypes. In this study we describe how this SNP array (encoded SolSTW array) was designed and analysed with fitTetra, software designed for autotetraploids. Genotypes from different countries and market segments, complemented with historic varieties and important progenitors, were genotyped. This comprehensive set of genotypes combined with the deliberate inclusion of a large proportion of SNPs with a low minor allele frequency allowed us to distinguish genetic variation contributed by introgression breeding. This "new" (post 1945) genetic variation is located on specific chromosomal regions and enables the identification of SNP markers linked to R-genes. In addition, when the genetic composition of modern varieties was compared with varieties released before 1945, it appears that 96 % of the genetic variants present in those ancestral varieties remains polymorphic in modern varieties. Hence, genetic erosion is almost absent in potato. Finally, we studied population genetic processes shaping the genetic composition of the modern European potato including drift, selection and founder effects. This resulted in the identification of major founders contributing to contemporary germplasm.

Introduction

The genetic diversity of cultivated potato, as studied today, is easier to interpret with insights from the past events that have shaped the gene pool. These past events include (1) the amount of genetic variation that was brought from South America to Europe since the late 16th century (Hawkes and Francisco-Ortega 1993), (2) the loss of diversity during the late blight epidemics in the 19th century, and more recently (3) introductions of (wild) Latin American species contributing to pathogen resistance, and (4) the more focussed breeding for specific niche markets. These genetic interventions leave their traces and can be recognized using modern DNA tools. Moreover, the maintenance of the original named varieties via clonal reproduction allows the historical analysis of the breeding process by comparing their genetic makeup with contemporary varieties.

A number of historic varieties dating back to the 19th century are still widely grown, as progress in variety improvement is limited due to the low reproduction rate and complex autotetraploid inheritance (2n = 4x = 48). Examples of historic varieties are Russet Burbank (1908), a Burbank (1876) mutant, the most important variety in the USA and Bintje (1910) ranking 1st in Belgium and 6th in the Netherlands in 2014. Throughout the 100 years of breeding hardly any increase in yield has been achieved (Douches et al. 1996), nevertheless major improvements have been made by introgression of resistance genes. Introgression breeding, practiced from the early 20th century onwards, focussed on late blight, cyst nematodes and viruses using *S. demissum, S. stoloniferum, S. tuberosum* Group Andigena clone CPC-1673 and *S. vernei* reviewed by Bradshaw and Ramsay (2005). This review describes the utilization of the Commonwealth Potato Collection (CPC) material, and similar work was performed in Germany and the Netherlands using other germplasm collections, e.g. Braunschweig Genetic Resource Collection (BGRC). Markers have been developed for the vast majority of the important resistance genes. Recent genepool-wide validation studies that aim to predict the presence of multiple resistances are still being performed with single locus marker techniques (Lopez-Pardo et al. 2013; Sharma et al. 2014). Unfortunately, it will take an additional effort to convert gel-based markers into SNP assays suitable for highly parallel genotyping methods.

In contrast to breeding for resistance, marker-assisted breeding for yield and tuber quality traits is still in its infancy. Marker-assisted breeding for such traits with a quantitative and polygenic nature will require a much deeper understanding of the loci and the beneficial and deleterious alleles involved. Genomic selection however does not rely on such information, but for both strategies the implementation of SNP arrays is urgent, to allow sufficient data collection to improve breeding efficiency. One of the first examples of highly parallel marker studies in potato made use of methods such as the Golden Gate assay (Anithakumari et al. 2010), the KASP SNP genotyping system (http://www.kbioscience.co.uk) (Lindhout et al. 2011) and more recently a SNP array with 8303 markers was developed (Felcher et al. 2012; Hamilton et al. 2011), which is currently

widely used in potato research (Hirsch et al. 2013; Manrique-Carpintero et al. 2014; Prashar et al. 2014).

Such a SNP array requires a SNP discovery study, which is facilitated by next-generation sequencing. Different approaches for SNP discovery are used such as (1) whole genome resequencing (Yamamoto et al. 2010), (2) transcriptome sequencing (Barbazuk et al. 2007; Hamilton et al. 2011, 2012; Trick et al. 2009) and (3) reduced representation sequencing based on restriction enzymes (Baird et al. 2008). These studies do not need any prior knowledge of a reference genome. The study of Uitdewilligen et al. (2013) used the potato reference genome (PGSC 2011) to perform a targeted resequencing of a subset of 800 genes (2.1 Mb) from the potato genome.

Many of these SNP discovery studies are based on a few genotypes, for example the parents of an important mapping population (Bundock et al. 2009), one rice variety compared to the reference genome (Yamamoto et al. 2010), six potato varieties (Hamilton et al. 2011) or four tomato varieties combined with two wild relatives (Sim et al. 2012). Even the most commonly used array in Arabidopsis was based on the sequence of only 19 accessions (Kim et al. 2007). This could result in an ascertainment bias when this array is applied on a much wider or different germplasm (Moragues et al. 2010; Thomson et al. 2012). In more recent studies larger SNP discovery panels are sequenced, for example in wheat (Wang et al. 2014). Also in potato a relatively large panel representative for the worldwide gene pool has been used for targeted resequencing (Uitdewilligen et al. 2013). These discovery studies on a wider gene pool are more suitable for the development of large arrays and can be applied to a much wider germplasm.

In this paper we describe the design of a potato 20K SNP array using the two major discovery studies available in potato (Uitdewilligen et al. 2013; Hamilton et al. 2011). On this SNP array a large number of relatively rare variants from Uitdewilligen et al. (2013) have been included. The array was used to genotype a comprehensive panel of 569 genotypes, which included many historically important varieties and progenitors, from different origin and market niches. We describe the analysis of this array with fitTetra (Voorrips et al. 2011) and subsequently we explore the (changes in the) genetic composition of the potato genepool. This allowed us to (1) identify introgression segments of different origin, (2) study the impact of breeding on allele frequencies in modern germplasm.

Materials and methods

Development of SolSTW array

A 20K SNP array was developed predominantly using a subset of the DNA sequence variants as described by Uitdewilligen et al. (2013). Several design criteria were taken into consideration to minimize the risk of assay failure due to flanking polymorphisms, and to maximize the ability to capture haplotypes across the diversity of potato germplasm. According to the manufacturer's instructions an Illumina SNP assay has to be free from flanking SNP over a region of preferably 60 bp at one side of the SNP to develop an optimal array. The SNPs selected from Uitdewilligen et al. (2013) were chosen to the following criteria (1) no InDels, tri- or quad- SNPs, (2) no InDels in flanking sequence, (3) only SNPs genotyped with a read depth ≥15× in at least 25 varieties, (4) minimum flanking SNPs free distance is five nucleotide positions, (5) maximum flanking SNPs at position 6–10 = 1, (6) maximum flanking SNPs at position 6–50 = 5 and (7) no Infinium type I assays. If both the left and right flanking sequences passed these criteria, then the side was chosen with the lowest number of SNPs in the first 10, or 25 or 50 bp. Next to criteria on technical suitability of SNPs, we applied genetic criteria to reduce redundancy of SNPs whilst maximizing the inclusion of SNPs from different haplotypes. To this end the genotyping calls across 83 varieties from Uitdewilligen et al. (2013) that were used to cluster SNPs with a Kendall tau test and correlated SNPs ($r^2 > 0.5$) were considered as one cluster. As a next step the clusters were ordered per gene and subsequently two SNPs per cluster per gene were selected. For the clusters without two SNPs within a gene, one SNP per cluster was selected. Finally singletons were added which were genotyped in at least 67 varieties, had a maximum of 2 flanking SNPs within 25 bp and no flanking SNPs within 10 bp. In this way SNPs are selected over the full length of large introgression segments and we tried to achieve a uniform distribution of SNPs across the length of the genome and across the depth of haplotypes. We did not filter SNPs according to allele frequency as calculated by Uitdewilligen et al. (2013), because our designed SNP array should allow to monitor the potential increase and decrease of both abundant as well as rare alleles during specific breeding efforts for specific market niches. We did not exclude variety-unique SNPs, with the exception of the excessive number of 2688 unique variants observed in variety Vitelotte Noire alone. This resulted in the selection of 15,123 SNPs from Uitdewilligen et al. (2013). Furthermore, 37 chloroplast markers were included (supplementary file 1).

Additionally, we included a subset of 4179 SNPs from the 8303 SolCAP array (Hamilton et al. 2011), which were reported to us to perform well on European tetraploid germplasm (Data not shown). To further improve genome coverage we analysed which PGSC superscaffolds of the potato genome were not yet or insufficiently represented. This resulted in the selection of an additional 284 markers from the 69,011 SolCAP SNPs discovered in Hamilton et al. (2011). Finally, we manually developed 124 SNPs

in functional genes involved in morphological and disease resistance traits. In Fig. 1 a Venn diagram is shown of the number of SNPs in each class mentioned above. The figure does not include 87 SNPs which have been found by both the SolCAP and our SNP discovery studies. This Venn diagram shows the attempted numbers of SNPs, but unfortunately, the total number of delivered SNPs was lower as shown in Table 1. To avoid any confusion this 20K SNP array is called the SolSTW array hereafter. It should be noted that Fig. 1 shows the attempted number of SNPs for the SolSTW array and the actual delivered number of SNPs for the SolCAP 8303 array. In supplementary file 1 additional information is specified for all markers as well as assay sequences.

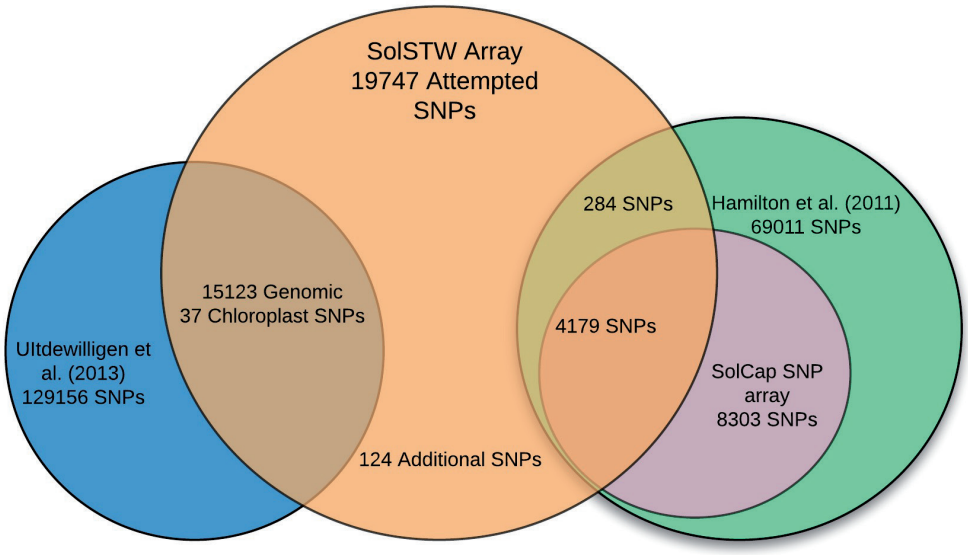

Fig. 1

SNP Resources. Venn diagram of SNPs used for development of the SolSTW array (*Orange*). SNPs originated from the 8303 SolCAP array are indicated in *purple*, SNPs originated from Uitdewilligen et al. (2013) are indicated in *blue* and SNP from Hamilton et al. (2011) are indicated in *green*. This figure does not display 87 SNPs included in the SolSTW array, which have been described by both the SNP discovery studies (Uitdewilligen et al. 2013 and Hamilton et al. 2011).

Plant materials

In this study we report about 639 DNA samples hybridized on the SolSTW array, of which 569 were unique genotypes consisting of 537 tetraploids and 32 diploids. To analyse the reproducibility, 39 tetraploid and 5 diploid samples were replicated on the array using DNA isolated from plants obtained from different sources. Besides these replicates the DNA from a single DNA isolation of clone RH89-039-16 (26 times) was used as an internal standard. The tetraploid genotypes consisted of 192 genotypes of a representative subset of commercial potato germplasm available worldwide, selected for the study of D'hoop et al. (2008) and complemented with a set of 173 advanced

breeding lines from Dutch potato breeders described in D'hoop et al. (2011, 2014). An additional set of 171 genotypes was collected, consisting of 51 varieties and 120 advanced breeding lines provided by Dr. Ronald Hutten (Wageningen UR Plant Breeding) and the company Meijer B.V. The names and additional information of the genotypes are provided in supplementary file 2.

Table 1

Numbers of attempted, delivered and successful SNP assays for the SolSTW array separated per SNP discovery study

Origin SNP	Attempted	Delivered	Successful	% Successful
PotVar SNPs (Uitdewilligen et al. 2013)	15123	13811	10707	77.5
Chloroplast SNPs (Uitdewilligen et al. 2013)	37	32	28	87.5
SolCAP 8303 Array (Felcher et al. 2012)	4179	3788	3561	94.0
SolCAP 69K detection (Hamilton et al. 2011)	284	246	202	82.1
Candidate genes	124	110	32	29.1
Total	19747	17978	14530	80.8

Data collection: DNA

Leaf material was collected for DNA extraction using the Thermo Scientific KingFisher Flex. DNA concentration was measured with the NanoDrop spectrophotometer and the DNA concentration was adjusted to ~50 ng µL–1 when possible. When DNA concentration was lower than 50 ng µL–1 the sample was still used up to a minimum of 25 ng µL–1 whilst for samples with lower concentrations a new DNA isolation was performed. For each 96-well plate the diploid genotype RH89-039-16 was included as a control. Infinium arrays were processed according to the manufacturer's suggested protocol at ServiceXS, Leiden, the Netherlands.

Genotype calling with fitTetra

For the genotype calling, fitTetra (Voorrips et al. 2011) and Illumina GenomeStudio software (version 2010.3, Illumina, San Diego, CA) were used. Whilst the polyploid module of GenomeStudio requires manual determination of the position and boundaries of the five clusters for each marker separately, fitTetra can perform this task fully automated. Therefore fitTetra was used to automatically score all markers. GenomeStudio was only used when the clustering of fitTetra resulted in inadequate genotype calling according to the criteria described below.

fitTetra first removes all data points with an overall R-value (overall intensity) below 0.2. Subsequently two default settings of fitTetra were adjusted (1) p.threshold was lowered from 0.99 to 0.95, which implies that there is 95 % confidence of a sample belonging to a cluster, resulting in less missing calls as compared to the more strict 0.99 threshold. (2) The peak.threshold was increased from 0.85 to 0.99, which allows SNPs with a very low

allele frequency (up to 99 % of all markers in 1 genotypic class) to be fitted by fitTetra. This was needed because the design of the array included a high number of low frequent SNPs. (3) the call.threshold was set to 0.60, resulting in the rejection of markers with more than 40 % missing values. Diploid samples are analysed along with the tetraploids to allow verification of the correct recognition of the nulliplex (AAAA), duplex (AABB) or quadruplex (BBBB) clusters.

Simultaneous analysis of diploids and tetraploids may however compromise Hardy–Weinberg assumptions implemented in fitTetra and this may result in the rejection of markers that display deviations from Hardy–Weinberg equilibrium. Therefore two runs were performed, one with and one without the diploid samples. Genotype calls of both fitTetra runs were compared and inspected for markers having obvious genotyping errors such as (1) a heterozygous genotype call for the reference genome genotype DM; (2) diploid genotype calls assigned to simplex or triplex clusters and (3) deviating Mendelian segregation in a tetraploid mapping population from a matching project (analysis of the tetraploid mapping population is beyond the scope of this paper). Markers showing one of these unexpected results could be the result of a poor marker or a poor clustering by fitTetra. The poorly clustered markers along with the chloroplast markers were manually scored with GenomeStudio. Additionally SNP markers initially rejected by fitTetra were visually inspected using the graphical output of fitTetra (Fig. 2) to diagnose the correctness of the rejection. SNP markers, rejected by fitTetra, but allowed manual scoring were re-joined with the final dataset using GenomeStudio.
Genotype calling with fitTetra results in dosage scores (0, 1, 2, 3 or 4) which reflects the Infinium assay design by the manufacturer, which uses the A nucleotide as a reference. Genotype calls were converted from the initial Illumina format into two derived datasets. The first dataset (used for association analysis) is based on the DM reference genome, where the SNP alleles are indicated as REF (DM) and ALT (non-DM). SNP dosage values ranging from 0 to 4 reflects the observed number of non-DM SNP alleles. The second dataset (used for population genetic analysis; this study) is based on the population minor allele frequency (MAF), where the SNP alleles are indicated as MIN (minor allele) or MAJ (major allele). SNP dosage values in this second dataset ranging from 0 to 4 reflect the observed dosage of the minor allele. This MAF dataset is more convenient, because none of the many haplotypes in the potato gene pool assumes a haplotype frequency exceeding 50 % (Uitdewilligen et al. 2013).

SNP dating to identify introgression breeding
In the second SNP dataset (MAF) the SNP dosage specifies the presence of specific SNP alleles. Subsequently we searched for the oldest variety that had this minor SNP allele. The age of a genotype is based on the year of market release as listed in the potato pedigree database (Van Berloo et al. 2007). For progenitors/unnamed genotypes the year

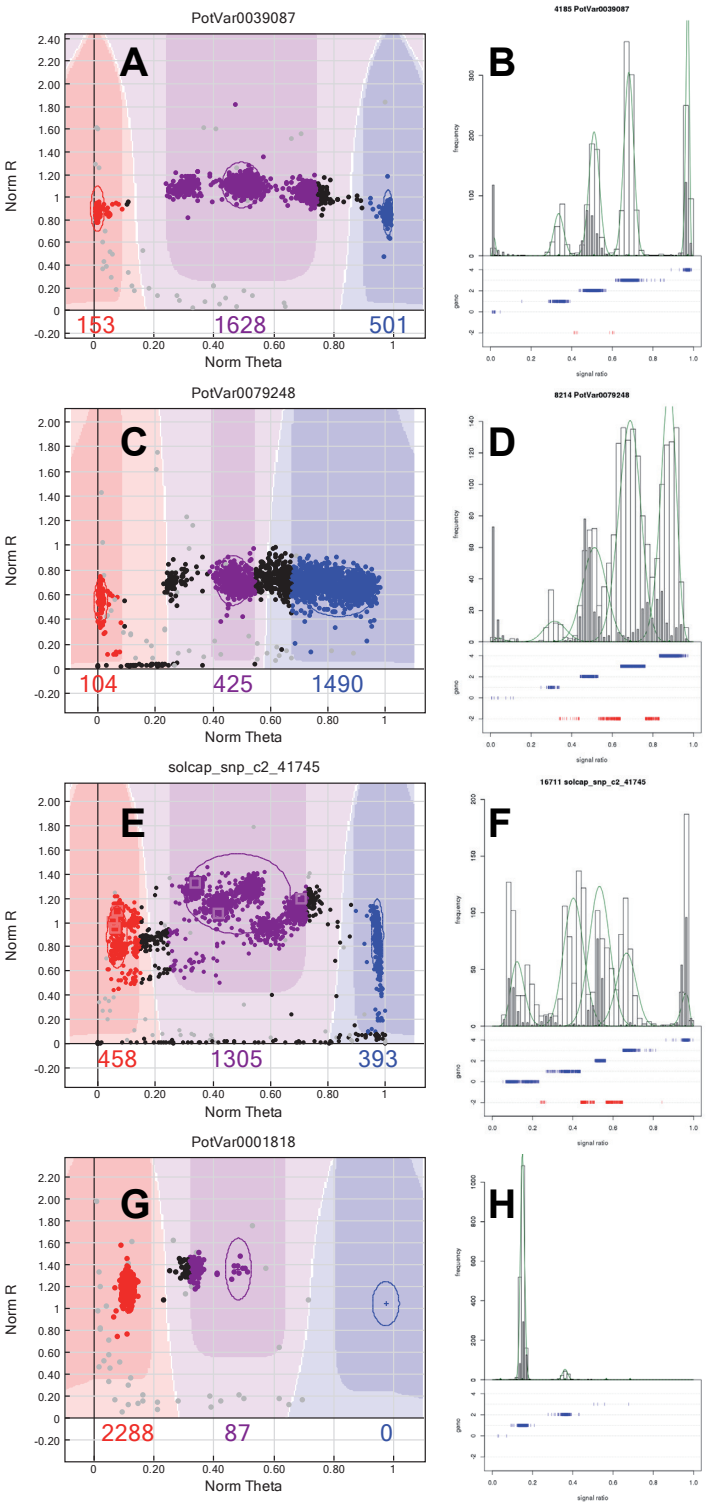

Fig. 2

GenomeStudio (*left*) and fitTetra (*right*) output. In panel **a** and **b** an ideal marker is visible, where five clusters are clearly distinguishable and the diploid samples (*grey bars*) cluster in the nulliplex, duplex and quadruplex clusters. In panel **c** and **d** a marker with a "cloud" of data points with overlap between clusters is visible, in panel **e** and **f** a marker with >5 clusters is visible and in panel **g** and **h** a marker with the nulliplex cluster shifted to the *right* is visible. *Grey bars* in fitTetra output represent diploid samples, these should cluster in three groups as in **a**. *Blue bars* in the bottom of fitTetra output represent the dosage calls and *red bars* represent genotypes which are in-between clusters

of crossing plus 10 years was taken. In this way the genotype and the year of introduction of each SNP was estimated. SNPs polymorphic in one of the 48 varieties released before 1945 are defined as "pre-1945" genetic variants. These SNP markers usually continue to be polymorphic after 1945. SNPs being monomorphic before 1945 and polymorphic in one or more of the 489 genotypes released after 1945 are assumed to be the result of introgression breeding and are defined as "new" or "post-1945" variants. The year 1945 was chosen because in 1946 Craigs Bounty was released, which is the oldest variety in our dataset harbouring wild species in its pedigree.

Analysis of changes in genetic composition of the potato gene pool

Changes in allele frequency were calculated to study the effect of breeding on the genetic composition of the gene pool. For this purpose the allele frequencies in the group of varieties released before 1945 ($N = 48$) was compared with the allele frequency in varieties released since 2005 ($N = 81$) and the allele frequency in the "starch" sub-populations ($N = 59$). Additionally the effect of allele dosage in important founders was compared with allele frequency changes. The subsets of genotypes are specified in supplementary file 2.

Results

Development of the SolSTW SNP array

The SolSTW SNP array combines SNPs from two discovery studies (Hamilton et al. 2011; Uitdewilligen et al. 2013). SNPs originating from Hamilton et al. (2011) were selected based on good performance in an earlier experiment using the SolCAP SNP array (data not shown) without any additional selection criteria. In contrast, the large set of 129,156 SNPs originating from Uitdewilligen et al. (2013) required stringent selection criteria since only a small subset could be selected. The high SNP density in potato allowed us to narrow down the number of potential SNP assays to 59,279 SNPs. Subsequently, redundancy amongst SNPs was reduced by clustering all SNPs according to SNP dosage as described by Uitdewilligen et al. (2013). This resulted in 7019 clusters and 5334 single SNPs (singletons). For around 5200 clusters, two or more SNP per cluster and gene were selected. Of the remaining approximately 1800 clusters, one SNP was selected and complemented with 2738 singletons, resulting in a total of 15,123 selected SNPs. SNPs originating from Uitdewilligen et al. (2013) will be referred to as PotVar SNPs. In Table 1 the attempted numbers of SNPs are shown.

Optimization of fitTetra with SolSTW array

Several runs were performed with fitTetra for genotype calling using the signal ratios obtained from the Infinium array. Over sequential runs, the programme settings were optimized and minor errors of the software were corrected. Two properties of the Infinium data initially resulted in erroneous clustering by the software. Firstly it appeared that the five clusters are not evenly distributed over the X-axis, as shown in Fig. 2a, b. In particular the three heterozygous clusters are closer to each other and relatively far from the two homozygous clusters. Secondly, the signal of the homozygous clusters is biased and not exactly at 0 or at 1 as shown in Fig. 2g, h. These modifications of the software are processed in the publically available version of fitTetra since autumn 2013 (https://www.wageningenur.nl/en/show/Software-fitTetra.htm)

Analysis of the SolSTW array with optimized fitTetra software

The improved version of fitTetra was used for the genotype calling of the SolSTW array. The genotype calling was performed twice, once using all genotypes and a second run without the diploid genotypes. The genotype calling without the diploid samples was used for further analysis, as inclusion of the diploid samples resulted in an additional rejection of 1184 markers, due to deviation from a Hardy–Weinberg test by fitTetra. The analysis of the tetraploid samples resulted in 15,271 fitted and 2716 rejected markers. Subsequently, a bi-parental tetraploid mapping population was used to identify SNPs where parental SNP dosage and offspring ratios were in disagreement. This is a putative indicator of poor SNP performance, and visual inspection of GenomeStudio

Table 2

Summary of total number of SNPs

	PotVar		SolCAP			
Chromosome	pre-1945	post-1945	pre-1945	post-1945	Total	% New
St4.03ch00[a]	42	7	45	1	95	13.1
St4.03ch01	902	478	405	17	1802	27.5
St4.03ch02	840	405	316	50	1611	28.1
St4.03ch03	662	307	293	45	1307	26.9
St4.03ch04	767	305	382	8	1462	21.4
St4.03ch05	830	254	214	24	1322	21.0
St4.03ch06	524	177	277	6	984	18.6
St4.03ch07	524	318	333	20	1195	28.2
St4.03ch08	485	154	334	4	977	16.2
St4.03ch09	535	206	302	9	1052	20.1
St4.03ch10	385	179	206	1	771	23.4
St4.03ch11	498	291	247	12	1048	28.8
St4.03ch12	479	185	190	22	876	23.6
Chloroplast	13	15	-	-	28	53.6
Total	7486	3281	3544	219	14530	24.1

Numbers of SNP markers per chromosome separated per origin (PotVar, SolCAP) and SNP age (pre-, or post-1945). Manually developed markers are within the set of PotVar markers

[a]St4.03ch00 lists marker that are located on unanchored scaffolds of the reference genome

output as shown in Fig. 2 resulted in the rejection of another 378 SNPs. In addition 1832 markers with a call rate below 95 % in fitTetra were visually inspected using GenomeStudio. The remainder of 6041 SNPs with good Mendelian fit and call rate >95 % were assumed to be good calls, and visual inspection was omitted. For the visual inspection fitTetra output was used as shown in Fig. 2b, d, f, h. In these figures diploid samples are illustrated with grey bars. The position on the X-axis of the diploids allows one to identify potentially poor markers, when diploid samples are in simplex or triplex clusters. As shown in Fig. 2d, f the diploid samples do not cluster together in the nulliplex, duplex or quadruplex clusters and therefore markers like these were removed. This incorrect clustering of diploids was predominantly observed in markers with more than 5 clusters as shown in Fig. 2e, f or markers with "clouds" of data points as shown in Fig. 2c, d. For 1206 of the 1832 markers with >5 % missing calls, visual inspection resulted in the removal of the markers from the final dataset. For 626 markers, fitTetra produced false negative genotype calls based on correct marker signal intensities. Such markers were manually re-scored using GenomeStudio. The 2716 rejected markers were visually inspected with fitTetra output as shown in Fig. 2, and scored manually if the

marker was mistakenly rejected. This resulted in the recovery of 843 markers. Of these 843 markers 689 had an allele frequency below 1 %, therefore these were correctly rejected based on the peak.threshold setting in fitTetra of 0.99. The remaining 154 were mistakenly rejected for unknown reasons.

Table 3

Assay failure as a function of chromosomal positions for PotVar SNPs

position on pseudomolecule	coding/non-coding	ok	Failed	Percentage
Euchromatin	Coding	4348	538	11.0
Border	Coding	313	46	12.8
Heterochromatin	Coding	221	42	16.0
Total (coding)		4882	626	11.4
Euchromatin	non-coding	3828	1214	24.1
Border	non-coding	318	216	40.4
Heterochromatin	non-coding	449	739	62.2
Total (coding + non-coding)		9477	2795	22.8

Number and percentage of successful and failed SNPs separated based on position on the pseudomolecules (Euchromatin, heterochromatin and border as defined by Sharma et al. 2013) and based on coding and non-coding regions

Reproducibility of genotype calls

As shown in Tables 1 and 2 the data collection with fitTetra and GenomeStudio resulted in a final dataset with a high number of 14,530 SNP markers. The genotype calls of the 39 replicated tetraploid samples showed a high concordance between replications. On average, only 3.3 calls (0.02 %) differed between the replicated samples of which 60 % are differences within the heterozygous clusters. Additionally for 74 (0.5 %) markers on average there was no call for either of the genotypes. The 26 replicates of the internal diploid control also showed highly concordant results. We observed seven markers with a deviating observation. In addition, we observed 66 markers with one or more missing calls, of which 50 % were caused by two of the twenty-eight replicates.

The percentage of missing calls was very low for the final dataset of 14,530 markers and 537 genotypes, with only an average of 95 missing calls per genotype and 3.5 missing calls per marker (0.65 %). For genotypes having wild species in their pedigree and not used in the SNP discovery panel of Uitdewilligen et al. (2013), the average number of missing values was much higher (184).

Analysis of factors influencing assay failure

Several possible factors that could cause assay failure have been examined. In Table 1 percentages of assay failure are shown based on the origin of the SNP assay. What

is clearly visible is that the SolCAP SNPs originating from the 8303 array are most successful (94.0 %), because these SNPs were tested before with the Infinium platform. The non-pre-tested SNPs from Hamilton et al. and the SNPs originating from the SNP discovery study of Uitdewilligen et al. (2013) show a lower percentage of successful assays (82.5 and 77.5 %, respectively). However, when considering markers in coding regions only, the assay failure rate of PotVar SNPs is much lower (11.4 %, Table 3). For SNPs that were manually developed the majority failed (70 %), this could be explained by the location in R-genes, which are members of a large highly variable gene family. In Table 3 percentages of assay failure of 12,272 SNPs are shown based on their localization in coding or non-coding regions, as well as based on their chromosomal position on the pseudomolecules (Sharma et al. 2013). The latter can be divided in euchromatin, pericentromeric heterochromatin and the border between the two. It is clear that SNPs localized in the pericentromeric heterochromatin are more likely to fail. However, more significant is the low percentage of assay failure in coding regions compared to non-coding regions.

The high nucleotide diversity of potato implies that SNP assays may be frequently affected by flanking SNPs. Therefore we aimed to target SNPs without flanking SNPs for assays, this is however problematic in potato due to its high SNP density. Consequently for many (34.8 %) SNP assays (originating from Uitdewilligen et al. 2013) on this array, known flanking SNPs are present. In Fig. 3a the percentage of assay failure of these PotVar SNPs is shown as a function of the distance of the flanking SNPs. This graph shows a trend where flanking SNP distance is correlated with assay failure. Additionally in Fig. 3b a correlation is shown between assay failure and the number of flanking SNPs. An increase in assay failure with more flanking SNPs can be observed. In addition the GC content was compared between successful and failed SNPs, however there was no significant relation between assay failure and GC content.

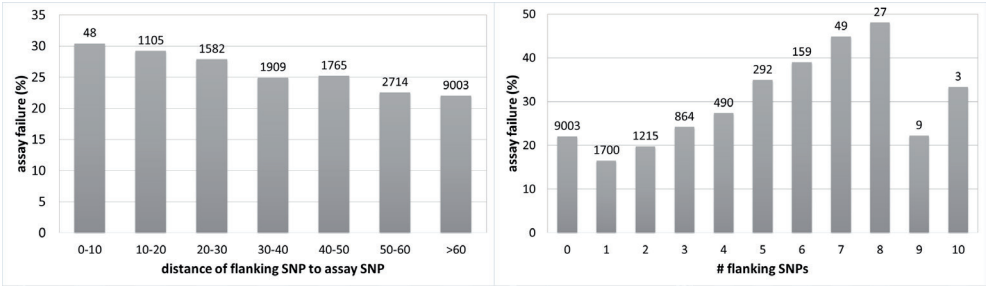

Fig. 3

Assay failure as function of flanking SNPs. (*Left panel*) Percentages of assay failure as function of the distance (in bp) of the first flanking SNP to the attempted SNP assay. (*Right Panel*) Percentage of assay failure as a function of the total number of SNPs observed in 50 bp flanking region.

Allele frequencies

The allele frequency distribution of SNPs across the 537 genotypes is shown in Fig. 4. PotVar SNPs, shown in the distribution (wide bars, left Y-axis) and SolCAP SNPs (narrow bars, right x-axis) differ greatly in allele frequency. PotVar SNPs are split in pre-1945 (dark blue) and post-1945 (green) SNPs. The average allele frequency of PotVar SNPs is 11 % and for SolCAP 22.7 %. This large difference in allele frequencies, also shown in Table 4, is not surprising since we deliberately did not exclude SNPs with a low allele frequency, clearly these were selected against in the design of the SolCAP array.

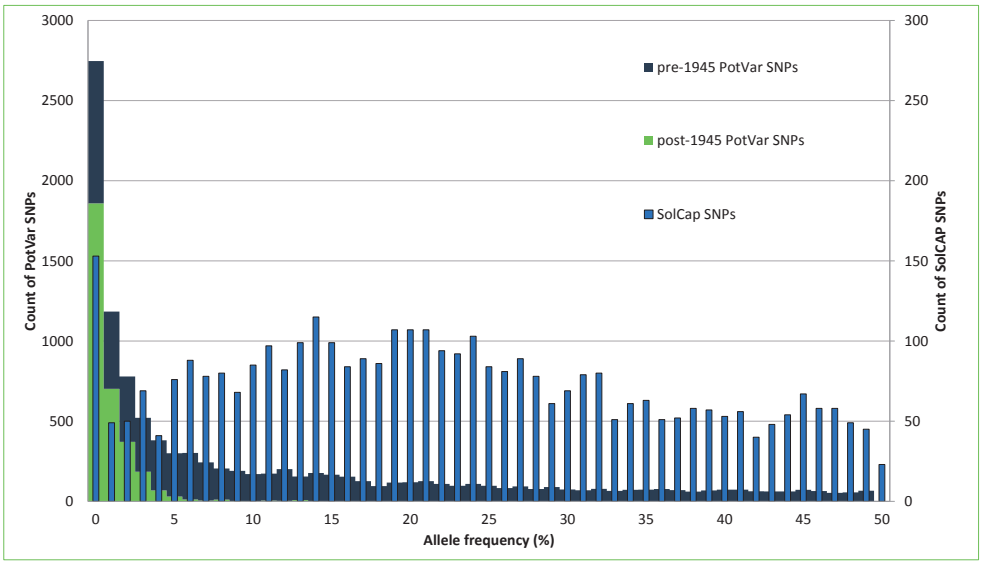

Fig. 4

Allele frequency distribution. Frequency distribution of Minor Allele Frequencies of SNPs in 537 tetraploid genotypes. The *wide bars* display the distribution of the 10,707 PotVar markers (Uitdewilligen et al. 2013), where the *blue* part indicates the proportion of pre-1945 SNPs and *green* the post-1945 SNPs. The distribution with *narrow* bars displays the Minor Allele Frequencies of the 3574 + 188 SolCAP markers (Hamilton et al. 2011). Chloroplast and manually developed markers are not included in this figure. The *left Y-axis* is for the PotVar SNPs and the *right Y-axis* is for the SolCAP SNPs

Identification of pre-1945 and post-1945 variation

The comprehensive sampling of the gene pool of cultivated potato allowed the evaluation of changes of the composition of the gene pool over time. This resulted in the identification of SNP markers, which are the result of introgression breeding and SNP markers that represent the initial genetic diversity within the founders of the contemporary gene pool. A SNP that is polymorphic in one of the 48 varieties released before 1945 is hereafter referred to as "pre-1945" SNP. This genetic variation most likely represents the material that was brought to Europe from the Americas between the 16th and the 19th century. A SNP marker that is monomorphic in one of the 48 old varieties, but polymorphic in more recent varieties/progenitors is hereafter referred to as "post-1945" variation. In Table 4 the large difference in allele frequency is visible between the post-1945 SNPs (average MAF = 1.4 %) and the pre-1945 SNPs (average MAF = 18.0 %). In Table 2 the numbers and percentages of post-1945 SNPs per chromosome are shown. In total 3500 (3281 PotVar + 219 SolCAP) SNPs are post-1945, which corresponds to 24.1 % of the SNP markers in this array. The detection study of Uitdewilligen et al. (2013) made a large contribution to this group of post-1945 SNPs (Table 2). The 219 post-1945 SNPs contributed by SolCAP are mostly introduced by variety Lenape (114 SNPs), of which two descendants (Atlantic and Snowden) were included in the discovery study of Hamilton et al. (2011). The chromosomal positions of post-1945 SNPs were analysed. It appears that post-1945 SNPs cluster together on chromosomes and in genotypes. In Fig. 5, a genome-wide plot is shown of the location of introgression segments first observed in six genotypes. Introgression segments differ greatly in size, ranging from very small (Y-66-13-636) to complete chromosomes (VTN 62-33-3). A nice example is the 97 SNPs first observed in Craigs Bounty (1946). This figure shows 95 SNPs in three introgression segments on chromosomes 5 (green), 10 (dark blue) and 12 (grey). Ten genotypes (VTN 62-33-3, Lenape, Mara, Urgenta, VE 71-105, AM 78-3704, Maris Piper, Craigs Bounty, Ulster Glade, VE 66-295) are responsible for the introduction of 50 % of post-1945 SNPs. A full table with numbers of SNP introduced per variety is shown in supplementary file 3.

Table 4

Numbers and average minor allele frequencies (MAF) of SNPs by age (polymorphic in pre-, post-1945 varieties) and discovery study [PotVar from Uitdewilligen et al. (2013), SolCAP from Hamilton et al. (2011)]

		Pre-1945 SNPs	Post-1945 SNPs	Total
PotVar SNPs	Numbers	7486	3281	10769
	Average MAF	15.2%	1.3%	11.0%
SolCAP SNPs	Numbers	3544	219	3763
	Average MAF	24.1%	1.5%	22.8%
Total	Numbers	11030	3500	14530
	Average MAF	18.0%	1.4%	14.0%

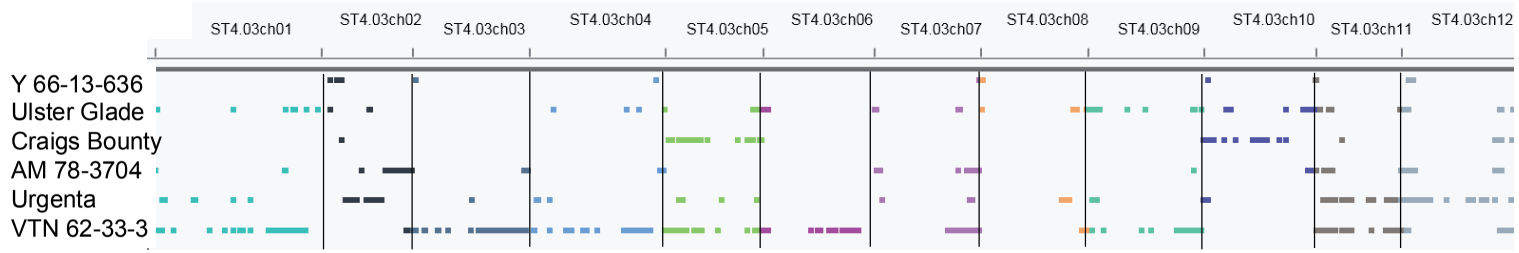

Fig. 5

Genomic position of newly introduced variation. Genome-wide plot of the coordinates of post-1945 SNPs on the DM reference genome where post-1945 SNP indicates the position of putative introgression segments first observed in six varieties. *Each dot* represents one SNP and it is visible that multiple varieties can introduce different haplotypes in the same region.

Processes that shape the genetic composition of the contemporary gene pool of potato

Several processes are shaping the contemporary gene pool of potato, such as the introduction of new genetic variants by introgression breeding. Introgressions cause the loss of existing variants by substitution. Selection will also influence the allele frequency, including breeding for specific market niches (e.g. starch varieties). In specific market niches, the limited gene pool is easily affected by random genetic drift (genetic erosion). These processes (introgression/substitution, selection, drift) were studied by comparing SNP allele frequencies between two groups. Firstly, the pre-1945 varieties were compared with the varieties released after 2005. Also, the pre-1945 varieties were compared against varieties from the "starch" subpopulation. For post-1945 SNPs significant increases of the allele frequency can be observed. In this study we analysed 246 varieties that were released between 1946 and 2005. In this group, 108 varieties contributed post-1945 SNPs, ranging from 1 to 447 post-1945 SNPs per variety (Supplementary file 2). From these 108 varieties 39 are shown in Fig. 6 and arranged in the order of market introduction. These 39 varieties are donors of those post-1945 SNPs that have attained the largest increase in allele frequency within the 242 varieties released after 2005. The negative slope perceived in Fig. 6 indicates that introgression segments introduced soon after 1945 could assume a higher allele frequency (up to 19 %) as compared to more recently introgressed haplotypes (up to 4 % increase). This suggests that a prolonged presence of a beneficial haplotype introgressed in the gene pool results in increasingly

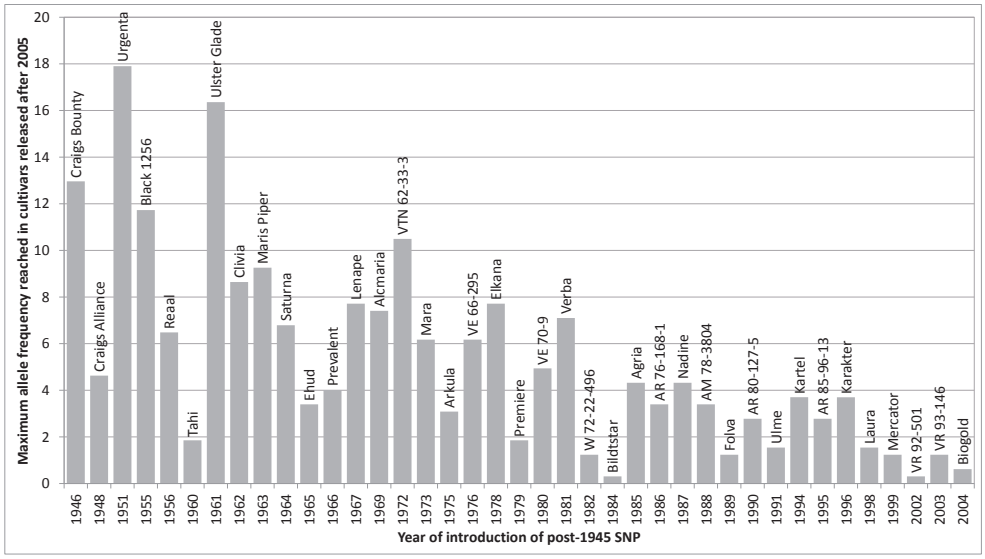

Fig. 6

Positive selection on newly introgressed SNPs. This figure shows the maximum allele frequency of post-1945 SNP reached in a panel of 244 genotypes released since 2005. The higher the *bar* the more frequent the SNP is observed in recent material, suggesting positive selection. The variety name above *each bar* indicates the first variety on the market showing a polymorphism and thus represents the founder genotype of the SNP

higher allele frequencies due to positive selection. Please note that a 4 % increase in allele frequency implies that almost 20 % of the varieties carry this haplotype in simplex condition, whereas a 19 % increase implies that more than half of the varieties are simplex or duplex and occasionally triplex.

In contrast, 50 % of all post-1945 SNPs remain below an allele frequency of 1 % and 549 SNPs were not polymorphic anymore (nulliplex) in varieties released after 2005. These 549 SNPs could be considered as lost, i.e. phased out soon after introduction. For the pre-1945 SNPs, 538 SNPs (4.9 %) were no longer polymorphic in contemporary varieties. These SNPs are also assumed to be lost during breeding. This may be due to selection, but random genetic drift is also plausible, because the initial allele frequency of these SNPs in old germplasm was already very low (1.4 % on average).

A comprehensive overview of the changes in allele frequency of all pre-1945 SNPs (in post-2005 and starch varieties compared with old varieties) is shown in Fig. 7. The largest column in the middle of the figure shows that the majority of the SNPs (6441 or 42 %) hardly changed in allele frequency during a century of potato breeding. Starch varieties show somewhat larger fluctuations in allele frequencies, because of an emphasis on introgression breeding for nematode resistance along with founder effects (discussed below). Figure 7 also suggests that larger numbers of SNPs have declined, as compared to the number of SNPs that show an increased allele frequency. This suggests that broadening of the genetic diversity by introgression since 1945 results in an

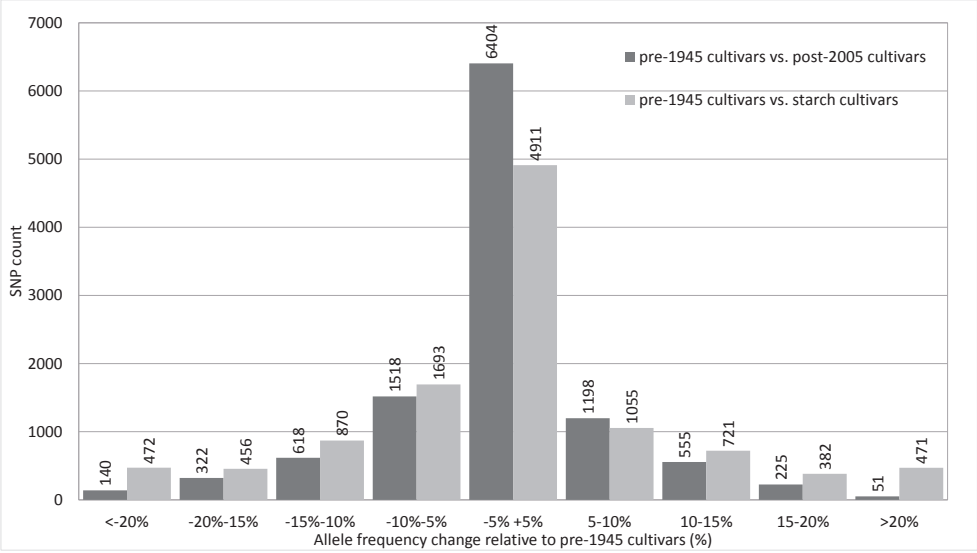

Fig. 7

Allele frequency changes. Distribution of allele frequency change of all pre-1945 SNPs is shown. The *dark grey bars* represent the number of SNPs and their change in minor allele frequency as compared between a panel of older varieties (market release before 1945) and a panel of new varieties (market release after 2005). The *light grey bars* show the comparison between older varieties and genotypes included in the subpopulation of starch varieties

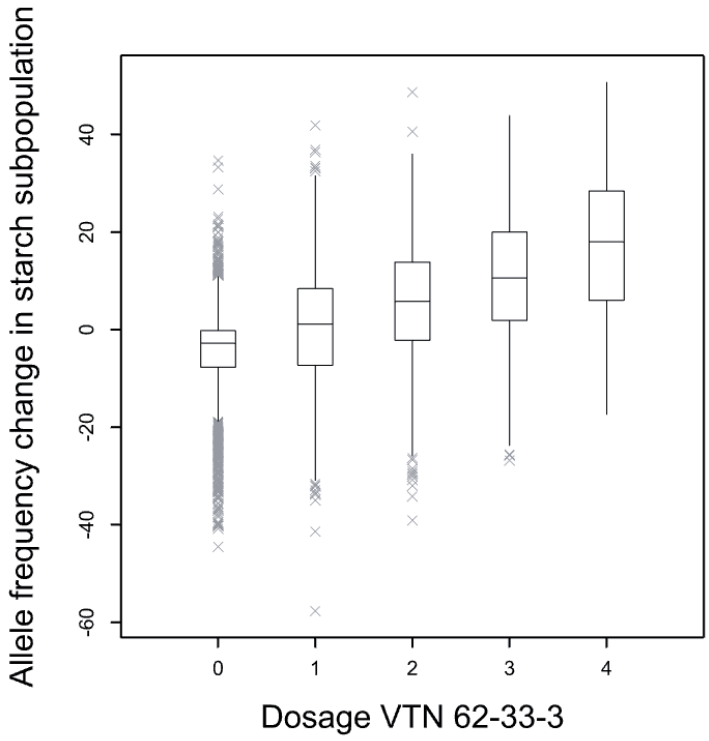

Fig. 8

Founder effect. The change of the Minor Allele Frequencies of pre-1945 SNPs within the subpopulation of starch varieties compared with pre-1945 varieties (*Y-axis*) as a function of the allele dosage of the minor allele in one of the most important progenitor clones (VTN 62-33-3). Here evidence is provided that a founder effect (initial allele dosage of a pre-1945 SNP) has a substantial impact, comparable to selection of a post-1945 SNP

overall net decrease of the frequency of pre-1945 haplotypes. In addition to these allele substitutions, founder effects may also reinforce this fluctuation. In Fig. 8 the change in allele frequency in the "starch" subpopulation is plotted against the allele dosage of an important progenitor (VTN 62-33-3). The figure clearly shows that a higher dosage of a SNP in an important founder contributes to the gain in allele frequency over time. The correlation between SNP dosage in a specific founder and allele frequency gain within the "starch" subpopulation was strongest for VTN 62-33-3 and AM 78-7804, two frequently used progenitors.

The processes underlying allele frequency changes over time: introgression, substitution, selection, drift and founder effects (frequent use of parents) are highly confounded. Still we assume that SNP variants that show the greatest increase in frequency are linked to important alleles for agronomic performance and vice versa. The most striking observations are that (1) 95.1 % of the pre-1945 SNPs are still polymorphic after 50–150 years of breeding and (2) we do not observe any fixation of pre-1945 SNPs in varieties released after 2005.

Discussion

Ascertainment bias

Here, we describe the development and evaluation of a SNP array for potato. Another SNP array named SolCAP is already widely used by the scientific community (Felcher et al. 2012; Hirsch et al. 2013; Lindqvist-Kreuze et al. 2014; Manrique-Carpintero et al. 2014; Prashar et al. 2014). We acknowledge the value of this array and reused 4179 SNPs with good assay performance. The other SNPs of this 20K SolSTW array were from a discovery panel of 83 tetraploids, comprising progenitors and varieties across breeding history, geography and market niche (Uitdewilligen et al. 2013).

Genetic diversity is unequally distributed across the gene pool. Therefore, a systematic deviation of a SNP discovery panel relative to the set of interrogated individuals will result in an ascertainment bias. Where SolCAP retrieved SNPs predominantly from North American varieties for the processing industry, the discovery panel of Uitdewilligen et al. (2013) contributed a relatively high number of SNPs typical for wild species introgression segments in progenitors. Although ascertainment bias is an important issue in the development and application of SNP arrays (Moragues et al. 2010; Thomson et al. 2012), it is difficult to quantify and difficult to avoid. A wider discovery study is in general better for SNP arrays intended for a wide range of applications. In view of the progenitor clones included in the discovery panel this array will be valuable to identify SNPs associated with resistance to cyst nematodes and viruses, but the array will be blind for East European haplotypes conferring e.g. resistance to Wart pathotype 18.

Assay quality

In view of the high SNP density in potato (Uitdewilligen et al. 2013) a comprehensive SNP discovery panel also allows the identification of flanking SNPs that could negatively influence SNP assay quality. Unavoidably, 34.8 % of the markers have a known SNP flanking the SNP assay and Figs. 2 and 3 illustrate their impact on assay quality. The analyses confirmed the negative effect of flanking SNPs, but also that their effect is proportional to the proximity and the amounts of flanking SNPs. Surprisingly, genomic position was a much stronger indicator for assay failure, where SNPs located outside exons or beyond gene-rich euchromatic regions were more likely to fail.

Data collection with fitTetra

In contrast to automated genotype calling in diploids, genotype calling in tetraploids is not a trivial job. Standard software such as GenomeStudio can handle tetraploid data, but it cannot automatically cluster the fluorescent signal into the five potential clusters. Manual scoring of each marker separately will give the best results, however due to the increasing size of SNP arrays the workload of manual scoring is prohibitively high.

Several methods have been proposed to automatically score tetraploid data. Hackett et al. (2013) used an algorithm to cluster ratios based on the genotype call of the parents in a segregating population. In the paper of Hirsch et al. (2013) a custom cluster file was generated within GenomeStudio. Here, prior knowledge of the clusters per marker is needed, since every marker produces a slightly different distribution of the five clusters. Since this was the first time the array was analysed, such a custom cluster file is not available. Consequently, we used software specifically designed for genotype calling of tetraploids, fitTetra (Voorrips et al. 2011) and were able to gain more experience with this software. During the analysis of our data a number of improvements in the fitTetra software could be implemented, as described in the results. We recommended users to always download the latest software version from our website (https://www.wageningenur.nl/en/show/Software-fitTetra.htm). With fitTetra we were able to score the dosage for 80.5 % of the markers. This is relatively high compared to 45 % in Hirsch et al. (2013) and 38 % in Lindqvist-Kreuze et al. (2014). Furthermore, the clustering by fitTetra appeared to be more accurate than genotype calling by GenomeStudio. The 0.02 % difference between replicated samples reported in our study compares well with the 1.7 % difference in genotype calls reported before (Hirsch et al. 2013). Nevertheless, we show that fitTetra does not always assign the correct genotype call to a cluster (Fig. 2g, h), or will erroneously cluster poor markers (Fig. 2c, f). Inclusion of diploid samples and a tetraploid bi-parental mapping population were extremely helpful to identify and discard poor SNP assays. Without these internal controls the quality of our dataset is expected to be much lower. An additional advantage of fitTetra is that the visual output of fitTetra is very helpful for identifying poor assays.

Dating of SNPs as a tool for reconstruction of the breeding history

A comprehensive sampling of genotypes of different ages has enabled us to assign a date to each SNP and to differentiate between "pre-1945" and "post-1945" genetic variants based on the year of market release. We observed that new genetic variants, cluster together in specific chromosomal regions and reside in specific genotypes (Fig. 5). For example, Craigs Bounty (released in 1946) is the oldest variety in our panel with introgressed chromosomal regions. In this variety, 97 SNPs are polymorphic, which were monomorphic in older pre-1945 varieties. These post-1945 SNPs most likely descend from a (S. commersonii x S. demissum) × (S. maglia × S. edinense) hybrid six meiosis back in the pedigree (Van Berloo et al. 2007) and originate from the work of Salaman (1985). Craigs Bounty is one of the first varieties with the R1 gene conferring late blight resistance from Solanum demissum on chromosome 5 (Toxopeus 1956). Therefore the SNPs on chromosome 5 are good candidates for tagging the haplotype containing the R1 gene. Subsequent linking of the pedigree to the first observation of CPC-1673 derived material (Maris Piper, 111 new SNPs) resulted in candidate SNPs on

chromosome 5 tagging the H1 resistance haplotype conferring resistance to Globodera rostochiensis.
The post-1945 SNPs, introduced with the market release of variety Lenape are most likely descending from the Solanum chacoense grand-grandparent. Why S. chacoense was used is not clear, but Love et al. (1998) describe Lenape as a first variety with a higher amount of solids, however it is also a variety with high glycoalkaloid content. Hence these SNPs could map nearby QTLs involved in dry matter content and/or glycoalkaloid content.
For most varieties that contributed post-1945 SNPs the source of introgressions could be deduced from pedigree information. However, our data suggest that the variety Urgenta introduced 178 post-1945 SNPs. This does not match pedigree information describing Urgenta as a pure "S. tuberosum" variety. Along with the observation that Desiree, a daughter of Urgenta, does not contain any of the introgression segments, we conclude that this sample was named Urgenta erroneously.

Processes that shape the genetic composition of the contemporary gene pool of potato
Few of the newly introgressed SNP alleles show a considerable increase in allele frequency within a subset of recent material (Fig. 6). Especially SNPs near the H1 and R3a/R3b loci, reach an allele frequency of 15 and 10 %, respectively. This example of positive selection for SNPs flanking the H1 locus can be explained by the need for varieties resistant to Globodera rostochiensis by potato growers. The increase in frequency of SNP alleles that belong to the R3a/R3b haplotype is not easily understood. This locus R3a/R3b was soon overcome by late blight and does not provide a detectable level of field resistance to Phytophthora infestans. Nevertheless we observed that a large region (5 Mb) was retained in more recent material. This suggests that other beneficial alleles linked to these R-genes are introduced in the potato gene pool that caused a positive selection on alleles in this region, which might be interpreted as linkage drag. In contrast, the majority of the post-1945 SNPs do not exceed an allele frequency higher than 1 %. Since this variation is not under positive selection we conclude that this variation is not adding anything to the potato genepool and it will be a matter of time that this variation gets extinct in the newly introduced varieties.

The loss of genetic variation is due to breeding andand has been described as genetic erosion. It is often thought that breeding will decrease the amount of genetic variation over time, also described as genetic erosion, described as genetic erosion. However, this assumed trend of declining diversity is not supported by molecular data (van de Wouw et al. 2010). To our knowledge this is the first study that used dated SNPs to compare the loss of old polymorphisms with the influx of new diversity due to introgression. In agreement with van de Wouw et al. (2010), we observed an insignificant amount of genetic erosion in potato. The limited numbers of SNPs not being polymorphic

anymore are most likely "lost" due to drift instead of selection against these SNP alleles. In fact the opposite is occurring. Whilst the majority of genetic variation that was present 100 years ago is still present in modern varieties, new genetic variation introduced in the last decades caused an increase of genetic variation in the potato gene pool. The lack of fixation of beneficial alleles supports the hypothesis that breeders select highly heterozygous offspring, allowing optimal heterosis. The tetraploid nature of potato prevents efficient selection against non-beneficial alleles and the net result is that genetic erosion scarcely takes place. There are major shifts in allele frequencies also described by Hirsch et al. (2013), however only a limited set of SNPs show this pattern (Fig. 7). These more substantial changes in allele frequency can result from selection but can also be explained as a founder effect, where a higher allele dosage for SNPs along with the frequent use of an important progenitor has impact on the change of allele frequencies in breeding (Fig. 8). The joint effect of selection and founder effects may easily explain an allele frequency change up to 50 %.

Future applications

This SNP array has been available for a short time, which is due to manufacturer's quality criteria for shelf life, amount of material synthesized and willingness to keep stocks. We do not regret this short availability and will not re-order the same array. Arguments to avoid repetitive use of the same array are the ever-changing gene pool and the ever-changing ascertainment bias if the SNP discover panel is at odds with the QTL mapping panel. Finally, technology is evolving at high speed. Sequencing costs are dropping and bioinformatics tools become more user friendly to arrive at more cost-effective sequencing-based genotyping strategies. For future applications supplementary Table 1, attached to this publication, offers a lasting resource of SNP loci that have been demonstrated successful. As shown by Felcher et al. (2012) and here the initial success rate of a SNP assay ranges between 40 and 70 %. This publication confirms that a SNP assay, once sufficiently tested, has a very high probability of being good forever. Indeed, the inclusion of SNPs that were tested before with the SolCAP array were re-applied. For this group of SNPs a very high success rate was achieved of 94 %. Whenever there is a need to generate fixed SNP arrays or KASP assays, it is recommended to tap from SNPs that have been demonstrated as successful before.

Author Contribution

Conceived and designed the experiments: PGV, HJvE, JGAMLU. Performed the experiments: PGV. Analysed the data: REV, PGV. Wrote the manuscript: PGV, HJvE. Edited the manuscript: HJvE, REV, RGFV.

Acknowledgements

We kindly acknowledge Wilbert van Workum and Marjolein Janssen from ServiceXS for their excellent service provide in the array hybridisation and data collection. The development of the array was financially supported by a grant from the Dutch technology foundation STW (project WPB-7926). PGV is supported by a grant of CBSG (Centre for BioSystems Genomics) and by potato breeding companies Agrico Research B.V., Averis Seeds B.V., HZPC B.V., KWS POTATO B.V. and Meijer B.V. We also thank Dr. Ronald Hutten for making a set of genotypes available.

Chapter 3

Evaluation of LD-decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato

Peter G. Vos[1,3], M. João Paulo[2,3], Roeland E. Voorrips[1], Richard G.F. Visser[1,3], Herman J. van Eck[1,3], Fred A. van Eeuwijk[2]

[1] Plant Breeding, Wageningen University & Research , P.O. Box 386, 6700 AJ Wageningen, The Netherlands
[2] Biometris, Wageningen University & Research, P.O. Box 16, 6700 AA Wageningen, The Netherlands
[3] Centre for BioSystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands

Published in Theoretical and Applied Genetics, First Online 3 Oktober 2016.

Abstract

The magnitude of linkage disequilibrium (LD) and its decay with genetic distance are important parameters in determining the resolution of association mapping for a given marker density. Insight in local and genome wide LD patterns is useful for assessing the desired numbers of SNPs on arrays. To study LD and LD-decay in tetraploid potato we simulated a panel of autotetraploid genotypes and used it to explore the dependence on: (1) the number of haplotypes in the breeding population or panel, i.e., the amount of genetic variation; (2) the percentage of haplotype specific SNPs (hs-SNPs). Several estimators for short range LD were explored such as the average r^2, median r^2, and other percentiles of r^2 (80%, 90% and 95%). For LD-decay, we looked at $LD_{1/2,90}$, the distance at which short range LD is halved when using the 90% percentile of r^2 as estimator for LD. Simulations showed that the relative performance of various estimators for LD-decay strongly depended on the number of haplotypes, although the real value of LD-decay was not influenced very much by this number. The LD-decay estimator $LD_{1/2,90}$ was chosen to evaluate LD-decay in a panel of 537 tetraploid potato varieties. When varieties of different age were compared, $LD_{1/2,90}$ values were 1.5 Mb for varieties released before 1945, and 0.6 Mb in varieties released after 2005. $LD_{1/2,90}$ values within three different subpopulations ranged from 0.7 to 0.9 Mb. $LD_{1/2,90}$ was 2.5Mb for introgressed regions, indicating that large haplotype blocks reside in the potato gene pool. In pericentromeric heterochromatin LD-decay was negligible. This study demonstrates that several related factors influencing LD-decay could be disentangled, and that the estimation of LD-decay has to be performed with great care and knowledge of the sampled material.

Introduction

Linkage disequilibrium (LD) is the non-random association between alleles at different loci in a breeding population, and can be estimated by the correlation between (SNP) markers. The amount of LD between loci is important for the success of forward genetic studies, such as Genome Wide Association Studies (GWAS), because the extent of LD determines the required number of SNP markers and the mapping resolution (Flint-Garcia et al. 2003). Association studies originated from human genetics (Hirschhorn and Daly 2005), where large designed bi-parental populations such as those used in plants, are impossible. The power of LD-mapping was soon recognized by plant geneticists and therewith extensive studies on LD were conducted in *Arabidopsis thaliana* (Nordborg et al. 2002; Kim et al. 2007) and in maize (Yan et al. 2009; Van Inghelandt et al. 2011).
In the classical sense the extent of LD at generation t, (D_t) is influenced by recombination frequency (r) between two loci and the number of generations (t) since the reference generation $t=0$, according to the formula $D_t = D_0(1-r)^t$. Factors as non-random mating, selection, mutation, migration or admixture, genetic drift, or a small effective population size, will all affect estimates of LD and LD-decay (Flint-Garcia et al. 2003).
In heterozygous outbreeders, such as potato, pairs of SNP alleles located on the same haplotype (linked in coupling phase) can display high values of LD and subsequent LD-decay is a function of the recombination frequency and the number of generations as described above. In contrast, LD between SNP alleles on different haplotypes (linked in repulsion phase) is not easily detected.
Self-fertilizing plants usually show less decay of LD, because the total number of effective recombination events in a homozygous genetic background is by far lower than the number of generations. Accordingly, LD is reported to decay at short distance (100-1500 bp) in an outcrossing crop species, such as maize (Remington et al. 2001; Tenaillon et al. 2001) and at large distance (up to 20 cM) in several selfing crop species such as barley (Kraakman et al. 2004) or durum wheat (Maccaferri et al. 2005) This is in contrast with natural populations of selfing species such as *Arabidopsis thaliana* (Nordborg et al. 2002) and *Medicago truncatula* (Branca et al. 2011){Branca, 2011 #35;Branca, 2011 #2;Branca, 2011 #2}, where LD-decay estimation suggests a much faster decay of LD (within 10kb for both species). Perennial or vegetatively propagated species such as potato and sugarcane have a long breeding cycle and therefore show a limited number of historical recombination events. Hence, LD decays relatively slow (Raboin et al. 2008; D'hoop et al. 2010) in spite of the outcrossing nature of these crops.
Various approaches exist to estimate LD and LD-decay. For LD, most approaches are based on the correlation calculated between marker pairs after giving numerical values to allele states, where r^2 or D' are commonly used. LD measures at short range are commonly used in an attempt to robustify LD estimates. Short range LD is calculated across a certain interval of genetic distances between marker pairs and then the mean

LD or a percentile of the LD can be used to define short range LD. Given a definition for LD, again various methods can be used to estimate LD-decay. Trend lines can be fitted based on LD measures as a function of genetic distance between markers, this can be done for the mean LD (Yan et al. 2009), the median LD (Myles et al. 2010), or an LD percentile (Adetunji et al. 2014). Additionally the mathematical function for the trend line to describe LD-decay can differ. A non-linear regression is most common (Delourme et al. 2013; Stich et al. 2013), but also more flexible functions as the LOESS function (Esteras et al. 2013) and a spline function (Zegeye et al. 2014) have been used. As an alternative to trend lines, thresholds can be defined at which LD stops to exist and we reach equilibrium, i.e, no effective correlation between alleles at different markers. The most commonly used threshold is an r^2 of 0.1, but a threshold of r^2 = 0.2 has been used as well (Delourme et al. 2013; Li et al. 2014). Adetunji et al. (2014) and Van Inghelandt et al. (2011) used a threshold based on background LD using the correlation between markers from different chromosomes. A further possibility to define an LD-decay measure is to identify the distance at which half of the maximum (short range) LD has decayed (Kim et al. 2007; Lam et al. 2010; Branca et al. 2011; Zhao et al. 2011). This value will be referred to as $LD_{1/2,90}$ in this study and describes the initial slope of the LD-decay curve. Combinations of trend line functions and LD-thresholds as described above result in differences in LD-decay estimates, which may severely hinder comparison between studies or species.

The gene pool of potato offers a unique opportunity to unravel the influence of various population genetic factors that affect genome-wide decay of LD. Many varieties have been kept alive by vegetative propagation, and our panel of 537 varieties includes both ancient and modern varieties. Additionally, a comprehensive pedigree database is available to know the number of generations between varieties of a finite gene pool comprising a few thousand varieties developed over at least two centuries (Van Berloo et al. 2007). We used 14,530 SNPs of which we know the physical and genetic position (Vos et al. 2015) and because every SNP is dated by the year of market release of the variety first showing the SNP variant, we can distinguish the recently introgressed haplotypes from the preceding ones.

In this study several estimators for LD-decay have been explored. To assist our evaluation of these estimators we also used simulated data, varying in number of haplotypes, i.e. the amount of genetic variation, and the percentage of haplotype specific SNPs (hs-SNPs). The performance of different LD estimators was compared (average, median, 80%, 90% and 95% percentiles for short range LD) as well as the performance for estimators of LD-decay. Special attention was given to estimators for the distance at which half of the short range LD decayed ($LD_{1/2,90}$). Subsequently, short range LD and LD-decay were evaluated in a panel of 537 tetraploid varieties, genotyped with a 20K SNP array (Vos et al. 2015). Within this set of genotypes LD-decay was estimated **(1)** in varieties of different age to study LD-decay over time, **(2)** within three subpopulations to study the

effect of population structure and **(3)** using SNPs that have their origin in introgression breeding (admixture). Furthermore, LD-decay was estimated using **(4)** SNPs with different minor allele frequencies (MAF) thresholds and **(5)** SNPs having different chromosomal positions (in pericentromeric heterochromatin and in euchromatin).

Materials and Methods

LD-decay estimators and LD-decay estimation

Pearson r^2 formed the basis for LD estimation. The correlations were calculated on SNP dosage (0-1-2-3-4), both for simulated SNP data as well as for the SNP array data from the variety panel. Short range LD was calculated based on markers pairs within 100kb for the variety panel and within 1cM for the simulated data. For short range LD five different estimators were used, (1) average of the correlation within the window, (2) 50% percentile (median), (3) 80% percentile, (4) 90% percentile and (5) 95% percentile. For both simulated data and real data all chromosomes were pooled in order to get a genome wide LD-decay estimation.

LD-decay was estimated by using a spline that was fitted on a chosen short range LD percentile using the RQSMOOTH procedure in GenStat. From the fitted spline, the distance at which half of the short range LD had decayed was calculated. Typically, the 90% percentile of the short range LD was used, $LD_{1/2,\ 90}$..

Background LD was estimated in the varieties panel using 50 randomly chosen markers per chromosome. With these 600 markers Pearson r^2 were calculated using all possible marker pairs from different chromosomes. The 95% percentile of all these pairwise correlations was used to estimate background LD.

Simulated data

To improve our understanding of LD decay and LD-decay estimators we simulated a series of tetraploid variety panels using PedigreeSim (Voorrips and Maliepaard 2012). Panels were simulated to resemble the European potato gene pool, as perceived by earlier SNP genotyping studies (Uitdewilligen et al. 2013; Vos et al. 2015). It is assumed that most of the genetic variation present in the contemporary cultivated *Solanum tuberosum* gene pool descends from a limited number of founders, and thus represents a limited set of founder haplotypes (Love 1999). In order to test the effect of the number of (founder) haplotypes on LD-decay either 6, 8, 10 or 12 founder haplotypes were simulated with allele frequencies ranging from 2.5 - 30%, according to a geometric distribution (Uitdewilligen, 2012) (**Table S1)**. Four haplotypes were randomly assigned to 10 tetraploid founder genotypes in generation 0. Initially the proportion of haplotype specific SNPs (hs-SNPs) was 100%. With such a simulated dataset we can monitor all recombinations, and their effect on LD-decay over time. Additionally we varied

the percentage of hs-SNPs from the initial 100% with four additional percentages of 75%, 50%, 25% and 0% hs-SNPs, by randomly assigning a SNP to multiple founder haplotypes. A decreasing fraction of hs-SNPs implies an overall increase of the minor allele frequency, because the sequence variant is no longer unique for one of the haplotypes but is shared by two or more of the 6, 8, 10 or 12 haplotypes. Hence, the fraction of hs-SNPs is confounded with average MAF.

The two variables, four different numbers of founder haplotypes and five different percentages of hs-SNPs, resulted in 20 simulated populations at generation zero, each composed of ten tetraploid individuals. The simulation of LD-decay involved eight generations of random mating, with 200 tetraploid offspring genotypes per generation using the PedigreeSim software (Voorrips and Maliepaard 2012) with eight chromosomes (replicates) of 50 cM each, with 501 SNP markers per chromosome separated by 0.1 cM and a centromere at the 20 cM position, with random chromosome pairing and a probability of 10% of quadrivalent formation vs. 90% of bivalent pair formation. The genotypic scores of the 8th generation were used for calculations on LD and LD-decay.

LD in a panel of 537 potato varieties

In addition to the simulated panels we evaluated LD and LD-decay in a panel of 537 tetraploid varieties and progenitor clones (**Supplementary file 1**). This panel was genotyped with a 20K SNP array (Vos et al. 2015). This data was analysed using different subsets of marker and/or genotypes in five experiments as described below In experiments 1 and 2 we make use of subsets of the genotypes and experiments 3, 4 and 5 make use of different subsets of markers.

1. **LD decay over time/generations**: It is known that LD decays over generations and distance. To study LD decay in time, the 537 genotypes were assigned to four groups according to year of market release. Group 1 contained genotypes released before 1945 (n = 45), group 2 contained genotypes released between 1945 and 1974 (n = 42), group 3 contained genotypes released between 1975 and 2004 (n = 195) and group 4 contained genotypes released after 2004 (n = 255). The age of a genotype is derived from the pedigree database (Van Berloo et al. 2007) where the year of market release was taken. For progenitor clones without market release we added 10 years to the year the cross was made, as perceived from the first two digits in the seedling code, because in general it takes ten years between making the cross and naming a variety.
2. **Population structure:** Population structure is one of the factors that may influence LD and likewise LD-decay estimates. For this purpose the population structure was estimated using all markers in STRUCTURE (Pritchard et al. 2000), with an analysis of K = 3 groups. Genotypes with a membership probability >0.5 in the STRUCTURE analysis for the "starch" subpopulation (N = 59) and the "Agria" subpopulation (N = 71) were analysed separately from the large "rest" group (N

= 407) containing all other genotypes. Additionally population structure was estimated with a principal coordinate analysis with 710 independent markers evenly distributed over the potato genome. Information of the grouping of genotypes is given in **supplementary file 1.** Group names are named as described before (D'hoop et al. 2008).

3. **Admixture:** In Vos et al. (2015) pre-1945 (old) and post-1945 (new) genetic variation was distinguished. Pre-1945 SNPs represent sequence variants that are polymorphic in varieties released before 1945. The majority of these SNPs continue to be polymorphic in more recent varieties. The new or post-1945 SNPs are monomorphic in old varieties and are therefore most likely the result of introgression breeding. Larger LD blocks are expected due to recent admixture with SNPs that originate from donor species. LD-decay was analysed separately for pre-1945 and post-1945 SNPs. Old varieties were removed from the latter analysis, because all new SNPs are monomorphic.
4. **MAF (minor allele frequency) thresholds:** LD-decay was analysed using SNPs with a MAF of 1.0 %, 2.5 % and 10 %, where MAF > 2.5 % is the default set of SNPs also used in experiment 1 & 2. Variation in MAF thresholds compares well with variation in the number of haplotypes and hs-SNPs in the simulation study to understand the effect of the amount of genetic variation on LD-decay estimates.
5. **Chromosomal position:** Recombination is suppressed in pericentromeric heterochromatin and should result in decreased LD-decay. Markers located in the pericentromeric heterochromatin, as defined by Sharma et al. (2013), were analysed separately from markers positioned on chromosomal arms (used in experiments 1 to 4).

The physical distance between markers was extracted from the SNP coordinates on the potato reference genome V4.03 (Sharma et al. 2013). Based on Sharma et al. (2013) we selected SNPs that were clearly located on the chromosomal arms. For experiments 1 & 2 a subset of 6133 markers was selected of which the minor SNP allele was present at least five times in each subgroup (age & subpopulation). The number of markers selected for each experiment on the variety panel is shown in **Table 1.**

Table 1

Number of SNP markers available in various datasets for experiments to evaluate LD-decay in a panel of 537 varieties

Experiment [a]	Exp. 1,2 and 4	**Exp. 3**	**Exp. 4**	**Exp. 4**	**Exp. 5**
Chromosome	**MAF ~> 2.5% [b]**	**Post-1945 SNPs [c] MAF > 1%**	**MAF > 10%**	**MAF > 1%**	**Pericentromeric SNPs, MAF ~> 2.5%**
St4.03ch01	780	145	694	969	139
St4.03ch02	773	62	625	957	190
St4.03ch03	489	52	445	644	111
St4.03ch04	575	138	477	787	124
St4.03ch05	548	40	459	793	142
St4.03ch06	470	59	418	586	96
St4.03ch07	550	46	466	639	117
St4.03ch08	456	73	384	585	69
St4.03ch09	424	87	382	568	88
St4.03ch10	316	47	272	401	64
St4.03ch11	435	61	388	592	125
St4.03ch12	317	76	272	405	89
Total	6133	886	5282	7926	1354

[a] Experiments 1-5 are described in the text.

[b] Each SNP marker is polymorphic in at least five individuals per age and/or structure group.

[c] The SNP markers are only polymorphic in genotypes released after 1945

Results

LD-Decay in simulated data

As mentioned before LD in tetraploid potato is mainly the result of physical linkage between two markers. We declare only the pairwise correlations that result from markers in coupling phase of interest for LD and LD-decay estimation. However, in contrast to diploids where phasing of haplotypes is feasible (Excoffier and Slatkin 1995), phasing information for tetraploid potato is typically lacking. Consequently LD estimation uses all pairwise allele combinations of marker pairs. This includes correlations between SNP alleles linked in coupling phase, but also less informative correlations between the SNP alleles linked in repulsion phase. To separate the informative (high LD-values as a result from linkage in coupling phase) from the non-informative (low LD-values as a result from linkage in repulsion phase) we have used simulated datasets with 100% haplotype specific SNPs with a known phasing of SNP alleles. Using such a dataset for conventional LD-decay estimation results in an LD-decay plot as shown in **Fig. 1a**. This LD-decay plot contains two kinds of pairwise correlations. Either there is a significant

correlation due to the initial linkage between two hs-SNP alleles in coupling phase, or there is immediate linkage equilibrium (LE) due to random chromatid assortment of alleles linked in repulsion phase (i.e. on different haplotypes). However, the known haplotype structure of these datasets allows us to separate the informative pairwise correlation between markers linked in coupling phase from less informative correlations between markers in repulsion phase, as shown in **Fig. 1b** and **Fig. 1c** respectively. The difference between gradual LD-decay as a function of genetic distance and immediate linkage equilibrium due to random chromatid assortment is obvious. **Fig. 1b** shows the fitted spline drops below the threshold of $r^2 = 0.1$ at a distance of ~13,5 cM. This distance of 13.5 cM appeared to be fairly constant across simulations with 100% hs-SNPs (**Fig. S1**). A second important observation is shown in the bottom row of **Fig. S2:** these graphs show that by adding more haplotypes in a simulated dataset also the percentage of non-informative (generally very low) LD-values increases, and therewith changing the estimation of LD-decay using the same estimator. This second observation is in conflict with the standard formula $D_t = D_0(1\text{-}r)^t$ where the factors t and r suggest an independence of LD-decay with the number of haplotypes in a population. Therefore we conclude that when we aim at estimating LD and LD-decay due to alleles at the same haplotype, the use of all allelic pairs causes a bias that is a function of the number of founder haplotypes, i.e., the genetic diversity.

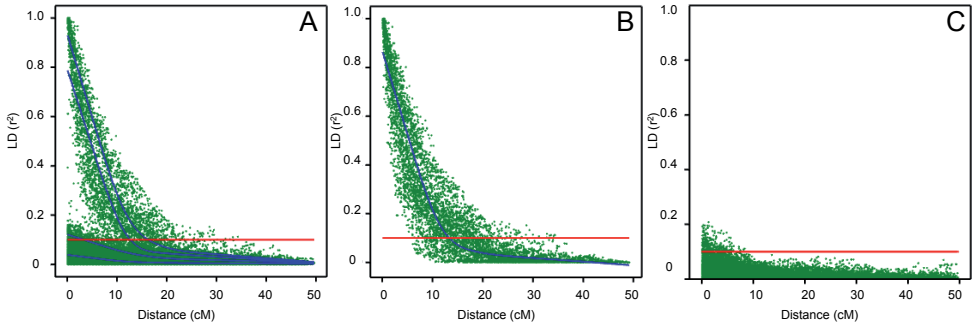

Fig. 1. LD-decay (in cM) in a simulated dataset after eight generations of random mating. 100% hs-SNPs and 6 founder haplotypes.

In panel (A) a traditional LD-decay plot is shown. The blue lines represent splines fitted for different percentiles (95%. 90%. 80%. 50%). from top to bottom. The LD threshold of 0.1 is indicated with a red line. Panel (B) is only based on pairwise correlations between marker alleles from the same haplotype. Here the blue line represents the 50% percentile. In panel (C) all pairwise correlations between marker alleles in linked in repulsion phase (different haplotypes) are shown.

Effect of the estimator on LD-decay estimates

In the simulated data, as an estimator for LD-decay we used the intersection of a 'significance' threshold ($r^2 = 0.1$) and a trend line. The trend lines shown in **Fig. 1a** and **Fig. S2** are based on spline fits on four LD percentiles (50%, 80%, 90% and 95%). Additionally LD-decay was calculated by the average r^2, and we looked at $D_{1/2,90}$ (based

on the 90% percentile). The LD-decay estimates differed greatly among simulated datasets and estimators (**Table 2**), and rarely reflected the expected value of 13.5 cM for haplotype sharing alleles as described in the previous paragraph. **Fig. 1a** shows that the 50% percentile (lowest blue line) is always below the threshold of r^2 = 0.1 and therefore here, and in many other simulations LD-decay could not be determined (shown as nd in **Table 2**). The average r^2, 50% and 80% percentile always result in a major underestimation of LD decay, while using the higher percentiles (90% and 95%) resulted in values closest to the earlier determined benchmark value of 13.5 cM. However, the 90% percentile also failed to estimate LD-decay accurately in the datasets with ten and twelve haplotypes. The simulations demonstrated that the number of haplotypes, or, amount of genetic variation had a strong effect on LD-decay estimates. We also show that use of the LD 90% percentile resulted in LD-decay estimates closest to 13.5 cM when six haplotypes were present and that the 95% percentile was optimal for 8 or 10 haplotypes. **Table 2** suggests that even a higher percentile should be used when 12 haplotypes are present, to compensate for underestimation of LD-decay.

Effect of the percentage hs-SNPs on LD-decay estimates

As explained in the previous paragraph the 90% percentile suits best for the situation with six haplotypes. The five datasets using six founder haplotypes result in fairly similar LD-decay estimates ranging between 12.3 and 12.8 cM. However, the simulated data demonstrate that the different LD-decay estimates are biased in their own way by the percentage of hs-SNPs. The average r^2 as well as the 95% percentile provide estimates suggesting a slightly faster LD-decay with fewer hs-SNPs. In contrast, the 50% and 80% percentile show an opposite bias. Remarkably, the values obtained with the $LD_{1/2,90}$ estimator do not seem to vary as much as the LD-decay values obtained with any of the other estimators. Therefore we propose that this estimator is a very promising estimator to compare LD decay across different studies. Unfortunately, the vast majority of LD studies do not yet use $LD_{½,90}$, and the outcome of $LD_{½,90}$ is difficult to compare with other LD-decay estimators.

Table 2 LD-decay estimates (in cM) from simulated data.

The values indicate the distances at which the five estimators for LD-decay (Average r^2 or a spline of four percentiles (50%, 80%, 90%, 95%) drop below the threshold for r^2 = 0.1. The values shown for $LD_{1/2,90}$ indicate the distances at which the 90% percentile decays to half of its initial value. Values in this table represent the average of 8 replicates (chromosomes). LD estimates that never exceeded the threshold for r^2 = 0.1 are not defined (shown as nd).

# Founder Haplotypes	% hs-SNPs	Average r^2 [a]	50% percentile [a]	80% percentile [a]	90% percentile [a]	95% percentile [a]	$LD_{1/2,90}$ [a]
6.0	100	5.7	nd	4.4	12.8	16.0	5.3
	75	5.5	nd	8.8	12.5	14.6	4.7
	50	5.2	nd	9.7	12.6	15.0	3.9
	25	4.8	2.0	9.7	12.3	14.4	4.7
	0	5.1	3.4	10.2	12.7	14.7	4.0
8.0	100	4.2	nd	nd	11.5	14.6	4.9
	75	3.2	nd	4.5	11.0	13.7	4.0
	50	3.2	nd	7.3	11.3	14.1	4.7
	25	2.6	nd	8.1	11.1	13.5	5.3
	0	2.4	0.4	8.8	11.5	13.6	5.0
10.0	100	2.7	nd	nd	3.4	13.9	4.6
	75	2.2	nd	0.8	9.2	13.1	4.1
	50	1.7	nd	5.1	9.9	12.9	5.1
	25	1.1	nd	6.6	10.0	12.2	5.1
	0	1.0	nd	7.6	10.5	12.7	5.5
12.0	100	1.3	nd	nd	nd	12.7	4.3
	75	0.6	nd	nd	7.1	12.0	4.1
	50	0.2	nd	2.8	8.2	11.3	5.4
	25	0.2	nd	5.0	9.1	11.8	5.4
	0	nd	nd	6.1	9.4	11.6	5.4

[a)] the standard errors of these estimates are very low (average = 0.25)

Table 3. Short range LD in simulated datasets.

Average and median of pairwise correlation (r^2) between pairs of markers within 1 cM.

# Haplotypes	% hs-SNPs	Average r^2	Median r^2
6 Haplotypes	100%	0.19	0.04
	75%	0.19	0.07
	50%	0.20	0.10
	25%	0.21	0.13
	0%	0.22	0.15
8 Haplotypes	100%	0.13	0.02
	75%	0.14	0.03
	50%	0.14	0.06
	25%	0.15	0.08
	0%	0.17	0.10
10 Haplotypes	100%	0.10	0.01
	75%	0.11	0.02
	50%	0.12	0.04
	25%	0.13	0.06
	0%	0.14	0.08
12 Haplotypes	100%	0.09	0.01
	75%	0.09	0.02
	50%	0.09	0.04
	25%	0.10	0.05
	0%	0.11	0.06

Short range LD in simulated datasets

We observed that $LD_{1/2,90}$ is the most constant LD-decay estimator in the simulated datasets. The $LD_{1/2,90}$ estimates rely on an estimation of the short range LD. The simulations show that short range LD estimates (within 1 cM) are also influenced by the number of haplotypes and hs-SNPs. Two remarkable correlations are shown in **Table 3.** First, the average pairwise SNP correlation is halved with a doubling of the number of haplotypes. Second, the median is decreasing significantly with an increase of the percentage of hs-SNPs. Based on these trends we can conclude that the amount of genetic variation can be approximated using the average of short range LD, and consequently the optimal percentile can be chosen to estimate LD-decay.

Short range LD in variety panel

Based on the empirical knowledge gained by preceding simulations it is important to select a suitable estimator for LD analysis in real data. For this purpose the short range LD is assessed using pairwise correlations between markers within 1kb, between 1kb and 10kb and between 10kb - 100kb. The average r^2 and 90% percentile were the highest in the subset of pairwise correlations of markers with 10kb to 100kb distance, suggesting no LD-decay within 100kb. Therefore pairwise correlations of markers within 100kb were used to estimate short range LD. Subsequently, the average r^2 and median r^2 of the short range LD were calculated for experiments 1 – 4, but not experiment 5 (LD in pericentromeric heterochromatin), and shown in **Table 4**.

The average r^2 ranged between 0.19 and 0.22 for the different age groups and structure groups of varieties (experiments 1 and 2, respectively). For experiment 3 we compared old and new (admixed) variation. The new variation resulted in a higher average r^2 indicating fewer haplotypes and a low median r^2 indicating a high percentage of haplotype specific SNPs. In experiment 4 we compared different MAF thresholds, which resulted in a lower average r^2 and median r^2 when more (lower frequent) SNPs were allowed. On average we found values around 0.2, which is a value similar to what we found in the simulated datasets with six haplotypes. In the simulated datasets the 90% percentile performed best, therefore we can conclude that the 90% percentile will result in a reliable estimate of LD-decay in the variety panel. This 90% percentile was subsequently used to describe LD-decay and to calculate the distance where the short range LD is decayed by 50% ($LD_{1/2,90}$).

LD-decay in different age groups, experiment 1

To evaluate how LD decays over generations within the potato genepool, the variety panel was divided in four age groups. Based on simulated data and short range LD in the variety panel a 90% percentile was used to describe LD-decay in the four age groups (**Fig. 2)**. LD-decay in individual chromosomes did not show significant differences and therefore all chromosomes were pooled. The group with the oldest varieties, released before 1945, displays the most LD (black curve), whereas the group with the youngest varieties, released after 2005, displays the least LD (blue curve). Remarkably, the different age groups decay to a different background level. Therefore, the intersection between the fitted spline and the threshold of $r^2 = 0.1$ results in large differences between the age groups, which might not represent the true LD-decay. Irrespective of unknown factors influencing background LD we observe that the slope of all curves flattens between a distance of 2 and 4 Mb. The $LD_{1/2,90}$ values (**Table 4**) of the different age groups (describing the slope of the first part of the LD-decay-curve) may represent the difference within the age groups better than the intersection of the spline with the threshold of $r^2 = 0.1$. The group with the older varieties reaches $LD_{1/2,90}$ at 1.5Mb and

Table 4. Short range LD in variety panel.

Average and median pairwise correlation (r^2) between marker pairs within 100kb in the different experiments on the variety panel. Additionally the 95% percentile of background LD (between chromosomes) and the distance at which half of the short range LD is decayed ($LD_{1/2,90}$) are shown for each experiment.

Estimator	Exp. 1. Age Groups				Exp. 2. Structure groups			Exp. 3	Exp. 4. Different MAF		
	<1945	1945-1974	1975-2005	Since 2005	Starch	Agria	Rest	New	MAF>1%	MAF (subgroups)	MAF>10%
Average r^2	0.22	0.22	0.20	0.19	0.20	0.21	0.20	0.28	0.14	0.20	0.24
Median r^2	0.11	0.11	0.09	0.09	0.09	0.10	0.09	0.01	0.03	0.09	0.13
Background LD *(r^2)*	0.13	0.14	0.04	0.03	0.10	0.08	0.02	0.07	0.02	0.02	0.02
$LD_{1/2,90}$ (Mb)	1.5	1.0	0.8	0.6	0.9	0.8	0.7	2.5	0.8	0.8	0.7

the group of young varieties reaches $LD_{1/2,90}$ at 0.6Mb (**Table 4**), suggesting that in 70 years of breeding a substantial reduction of LD has been achieved.

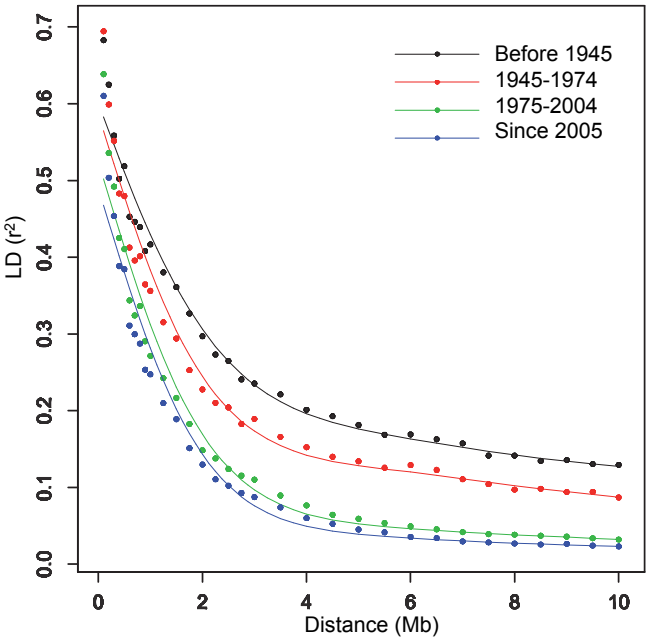

Fig. 2 LD-decay over generations.

The 90% percentile LD-decay splines are shown for varieties from four different age groups.

LD-decay in different structure groups, experiment 2

In order to estimate the effect of population structure on LD-decay, we analysed LD-decay within subgroups. **Fig. 3a** shows a principal coordinate (PCO) plot, where the colour of each variety represents a group membership as identified with STRUCTURE (Pritchard et al. 2000). The concordance between the analyses of PCO and STRUCTURE is high, even though the first two dimensions of the PCO explain only 3.9% and 2.8% of the variation. The red squares identify modern varieties selected for processing industry, and related to the variety Agria. The blue circles identify starch varieties and the green triangles represent all other varieties, mainly fresh consumption. In the principal coordinate and STRUCTURE analyses additional groups have been considered. The large green group could be separated in a third group with contemporary varieties and a fourth group representing heirloom varieties. The PCO axis separating the contemporary and heirloom varieties explained less than 1% and was therefore not used in this experiment. **Fig. 3b** shows the spline of the 90% percentile for these three different subpopulations. Again the curves flatten to different background levels between 2 and 4 Mb. However,

the initial slope of these curves show less difference compared to the age classes resulting in $LD_{1/2,90}$ values ranging between 0.7 Mb for the "rest" group and 0.9 Mb for the "starch" group. Selection for specific market niches resulted in a small reduction of LD-decay (**Fig. 3b**).

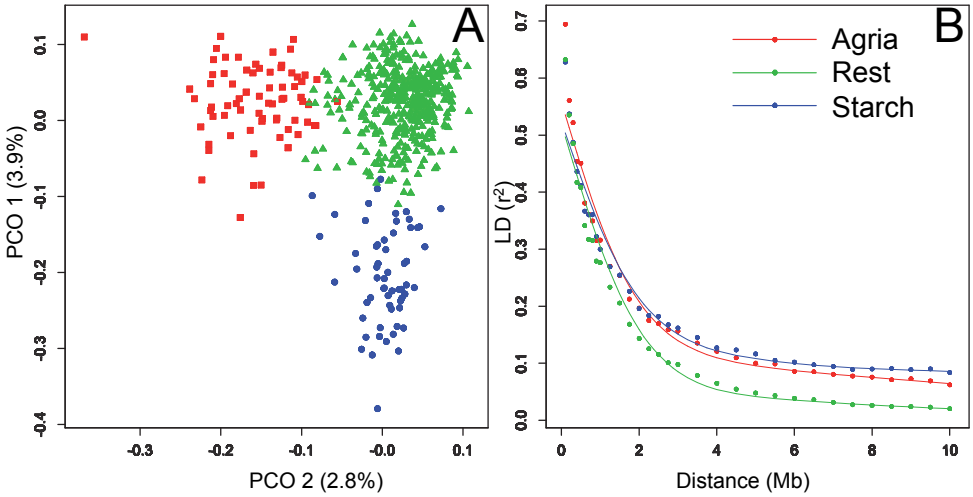

Fig. 3 LD-decay and population structure.

A: Principal coordinate analysis showing varieties. based on >0.5 group membership probability obtained by STRUCTURE. **B**: The 90% percentiles show LD-decay within subgroups of varieties. Red squares represent modern processing varieties related to Agria. Green triangles represent conventional varieties. Blue circles represent modern varieties bred for starch industry.

The effect of admixture, MAF-threshold and chromosomal position on LD-decay (experiments 3, 4 and 5)

New sequence variants have entered the potato gene pool due to introgression breeding since 1945 (Vos et al, 2015). The consequences of admixture between wild and elite material on LD decay could be analysed by comparing LD decay among SNPs that are polymorphic in material released before or only after 1945. **Fig. 4** shows the reduced LD decay perceived between SNPs on introgressed haploblocks (blue curve), that introgressed trait variation (e.g. resistance genes) could be detected with SNPs at several Mb distance, due to large haploblocks.

The black, turquoise and red curves in **Fig. 4** represent LD-decay based on subsets of SNPs with different minor allele frequency thresholds, where inclusion of more infrequent SNP alleles seemingly results in faster LD-decay (black curve) as compared to a more stringent threshold (MAF>10%, red curve). These curves drop below an r^2 threshold of 0.1 at significantly different physical distances. In contrast, the $D_{1/2}$ estimates shown in **Table 4** are remarkably similar and suggest a decay of LD at distances ranging from 0.7 and 0.8 Mb.

No decay of LD up to 10Mb was observed between SNPs at physical coordinates that belong to the pericentromeric heterochromatin (**Fig. 4** green curve), because of suppression of recombination in centromeric regions.

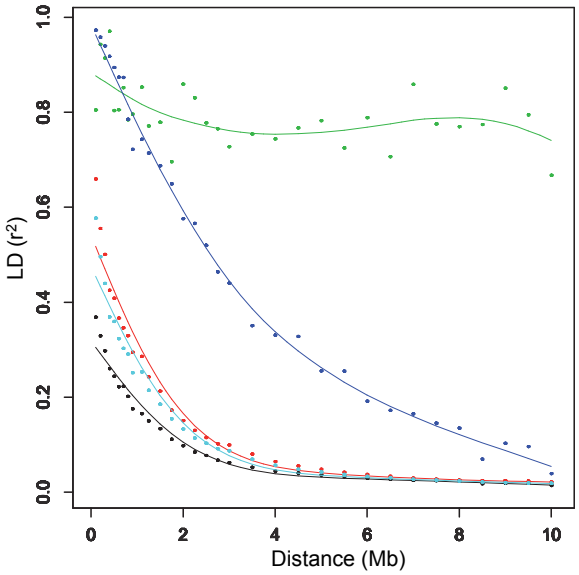

Fig. 4 LD-decay perceived with different subsets of SNPs.

The blue curve represents the LD-decay between SNPs on haploblocks introgressed since 1945. The red. turquoise and black curves represent LD decay between subsets of SNPs with different MAF thresholds. Red = MAF>10%. Turquoise is MAF>2.5%. black is MAF>1%. The green curve represents LD-decay between SNPs with physical coordinates located in the pericentromeric heterochromatin.

Discussion

Estimation of LD and LD-decay is a highly complicated matter without obvious consensus in literature on the preferred approach. LD-decay is influenced by many factors, which usually cannot be analysed separately. In this study we used simulated and real data representing tetraploid germplasm to understand which factors affect LD and LD-decay estimation. By doing so we gained more insight in LD patterns in potato and its implications for GWAS and the design of genotyping arrays.

In general the justification of the number of SNPs on the SNP-array is calculated by dividing the total length of a map/genome by the distance where a genome wide estimate of LD reaches threshold (often $r^2 = 0.1$). In the results we show that at ~2 Mb all curves start to flatten. When assuming a genome size of 400 Mb (including the gene rich arms and excluding the pericentromeric heterochromatin (Sharma et al. 2013) we can divide this 400 Mb by 2 Mb, suggesting that 200 markers are required per haploid genome

to detect QTLs. However, one could argue that an r^2 of 0.1 is too low for genome wide QTL discovery. On the other hand we show that within 100 Kb no LD-decay is observed, and consequently 4000 markers are needed to cover one haploid genome. In addition to the physical length of haploblocks also the number of haplotypes per locus is at stake to determine the number of SNPs on a SNP-array. The number (minimum of 200 and maximum of 4000) has to be multiplied by the number of haplotypes in a population. No accurate estimates are known on the number of founder haplotypes in the potato gene pool, but when assuming an average of 10 haplotypes, then 40,000 SNPs is an upper bound for QTL discovery. The 20K SolSTW array, with 14,530 SNPs does not reach this upper bound, but should still be able to detect sufficient QTLs.

Simulated data

The main conclusion drawn from simulated data is that estimates for genome wide values for LD-decay depend strongly on the estimator. Up to 10-fold underestimations of LD-decay were observed (**Table 2**), which has major implications for the required number of SNP markers for a GWAS, as well as the interpretation of the size of a candidate gene region. A second outcome of this study is the appreciation of short-range LD values to gain insight in the amount of genetic diversity.

Simulations showed that the average short range LD values halved when the amount of genetic variation was doubled, because of the decreasing signal of linkage disequilibrium from marker-pairs in coupling phase. This is caused by the fact that only SNPs residing on the same founder haplotype will result in high pairwise correlations. Correlation between markers on different founder haplotypes will always result in low correlation. Consequently the percentage of high pairwise correlations will decrease with the introduction of more haplotypes.

Short range LD in the variety panel

In the literature only relatively high averages of short range LD values have been reported, for example in crop species such as wheat (Würschum et al. 2013)and maize (Yan et al. 2009), where average initial LD of $r^2 = 0.32$ and $r^2 = 0.24$ respectively were observed, in (Wang et al. 2013) even averages of $r^2 = 0.5$ are observed. We observed a lower average r^2 for short range LD in our data (r^2 is between 0.19 and 0.22). This suggests that in these studies either only a limited amount of genetic variation is sampled or these studies dealt with more ascertainment bias compared to what we sampled in potato.

We propose that an estimator for LD-decay should be unbiased for, or adapted to the amount of variation present in the gene pool, to allow interpretable comparisons across species. For this purpose the average of the short range LD within the variety panel was compared with short range LD in simulated data. The observed average r^2 of approximately 0.2 in the subgroups corresponds well to values observed in simulated data with six founder haplotypes. Therefore the 90% percentile is most suitable for

analysing LD-decay. Indeed, this number of haplotypes is within the range of haplotypes found in the potato germplasm. Earlier haplotyping studies showed five haplotypes of the *LCYe* gene to 16 haplotypes of the *GWD* gene (Wolters et al. 2010; Uitdewilligen 2012).
Different background levels of LD makes it very difficult to determine one threshold at which linkage equilibrium is reached, therefore we focused on the initial slope of the LD curve and used the $LD_{1/2,90}$ values to compare LD-decay within several subsets as described in experiments 1 to 5.

LD-decay in age classes
To study how LD has decayed over the last century we compared old and recent varieties. We observed a decrease in LD over the last century from a $D_{1/2}$ of 1.5 Mb in old varieties to 0.6 Mb in recent varieties, suggesting that haploblocks still have a considerable length. Long haploblocks reflect a breeding history with typically a few meiosis in a century of potato breeding, where newly introduced varieties can be as little as six meiosis away from an ancestral variety from the 19th century (Van Berloo et al. 2007). In sexually reproducing crops and natural population many more meioses take place annually, and therefore one can imagine that haploblocks in potato stretches much further than in sexually propagated outbreeders. Van Inghelandt et al. (2011) also performed an analysis between old and new genotypes, however they observe an increase in LD in more recent material due to fixation for favourable alleles.

Population structure and LD-Decay in structure groups
Population structure is a confounding factor, influencing the associations in association mapping, resulting into false positive associations. Therefore it is essential to understand the population structure. We observed a weak population structure with PCO1 and PCO2 only explaining 3.9% and 2.8% respectively, similar to earlier potato studies (D'hoop et al. 2010; Uitdewilligen et al. 2013). Other studies (Li et al. 2010; Fischer et al. 2013; Stich et al. 2013) report on the absence of significant population structure. The difference between these studies could be explained by the sampling of the Dutch germplasm, where structure groups may result from Dutch breeding efforts. The structure group "starch" is mainly the result from specific breeding of high starch potato varieties within one breeding company. The second group is caused by the frequent use of the variety Agria as parent. Almost every variety within this group has Agria as parent or grandparent and these varieties have all been bred for the processing industry. A higher background LD was observed within these subgroups, as compared to the "rest" group, which could be the result of population structure. LD dropped below the traditional threshold of 0.1 at longer distance within these structure groups, as previously shown by D'hoop et al. (2010). However, the $LD_{1/2,90}$-values showed stable haploblock lengths, ranging from 0.7 Mb to 0.9 Mb.

Reduced decay of LD due to admixture

Vos et al. (2015) argue that the genetic variants within the potato germplasm can be divided into groups of SNPs predating 1945 and post-1945 variation, based on the year of market introduction of the variety. In this study LD decay was estimated using pre-1945 and post-1945 SNPs separately. The reduced decay of LD among post-1945 SNPs implies that introgressed haplotypes are substantially longer compared to haplotypes from earlier varieties. Here we have implicitly defined the length of a haplotype as the physical size of a genomic region flanked by historical recombination events. The dating of SNPs allowed us to quantify the effect of admixture on LD-decay. The data suggests that within contemporary varieties the size of haploblocks is highly variable.

Effect of MAF on LD-decay

In many studies a restriction on the MAF is applied, where a 5% cutoff is most commonly used (Zhao et al. 2011; Delourme et al. 2013; Esteras et al. 2013; Wang et al. 2013; Würschum et al. 2013; Adetunji et al. 2014; Li et al. 2014). In some cases a 10% (Hyten et al. 2007; Comadran et al. 2011) or even a 20% cut-off (Branca et al. 2011) is used. In this study we showed that a restriction on the MAF significantly reduces the average r^2 and therewith influences the LD-decay estimation when the intersection of a trend line and a threshold is used (**Fig. 4**). The effect of MAF thresholds was previously shown by (Yan et al. 2009). However, the $LD_{1/2,90}$ values (**Table 4**) were hardly affected by MAF thresholds.

Final remarks

Our analyses show that different estimators of LD and LD-decay can be chosen, and this choice will result in different estimates of LD-decay. In general we conclude that the $LD_{1/2,90}$ value offers the most consistent estimates of LD-decay and performed best in our study. Only a few studies use this estimator (Kim et al. 2007; Lam et al. 2010; Branca et al. 2011; Zhao et al. 2011) and justify a comparative analysis across species. In potato the distance where half of the initial LD is decayed is at least 600 Kb which is substantially longer than values observed in rice (100-300Kb (Zhao et al. 2011) or 3-4Kb in *Arabidopsis thaliana* (Kim et al. 2007) and *Medicago trunctula* (Branca et al. 2011). Unfortunately, no earlier study in potato used the $LD_{1/2,90}$ estimator, preventing us to compare our data with previous estimates of LD-decay in potato. On the other hand a general trend is that background levels of LD are reached at a distance between 2 and 4 Mb. This distance is equivalent to a genetic distance of 5-10 cM, which is in agreement with the 5 cM reported by (D'hoop et al. 2010) and the 10 cM reported by Simko et al. (2004). The remarkable low value of LD decay in 275 bp physical distance as reported by Stich et al. (2013) can now be understood as the consequence of the choice for the an LD-decay estimator using the average r^2 in combination with a non-linear regression.

Author contribution statement

Conceived and designed the experiments: PGV, HJvE, FAvE. Performed the experiments: PGV, REV. Analysed the data: PGV, MJP, FAvE. Wrote the manuscript: PGV, HJvE. Edited the manuscript: PGV, HJvE, FAvE, RGFV.

Acknowledgments

The collection of SNP data was financially supported by a grant from the Dutch technology foundation STW (project WPB-7926). PGV is supported by a grant of CBSG (Centre for BioSystems Genomics) and by potato breeding companies Agrico Research B.V., Averis Seeds B.V., HZPC B.V., KWS POTATO B.V. and Meijer B.V.

Table S1.

Allele frequencies of the founder haplotypes as used in simulated tetraploid potato genotypes.

6 Haplotypes	8 Haplotypes	10 Haplotypes	12 Haplotypes
0.3	0.25	0.25	0.175
0.25	0.175	0.15	0.15
0.175	0.175	0.15	0.125
0.125	0.15	0.125	0.1
0.1	0.1	0.1	0.1
0.05	0.075	0.075	0.075
	0.05	0.05	0.075
	0.025	0.05	0.05
		0.025	0.05
		0.025	0.05
			0.025
			0.025

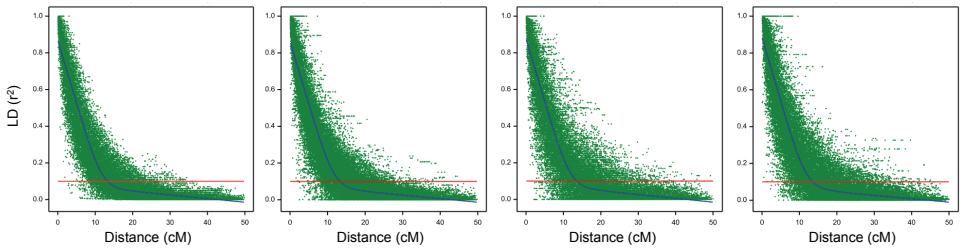

Fig. S1.

LD-decay curves from simulated data with 100% haplotype specific SNPs with 6 haplotypes (left) to 12 haplotypes (right). Only the pairwise correlations are shown resulting from markers that were linked in coupling phase in the founder genotypes.

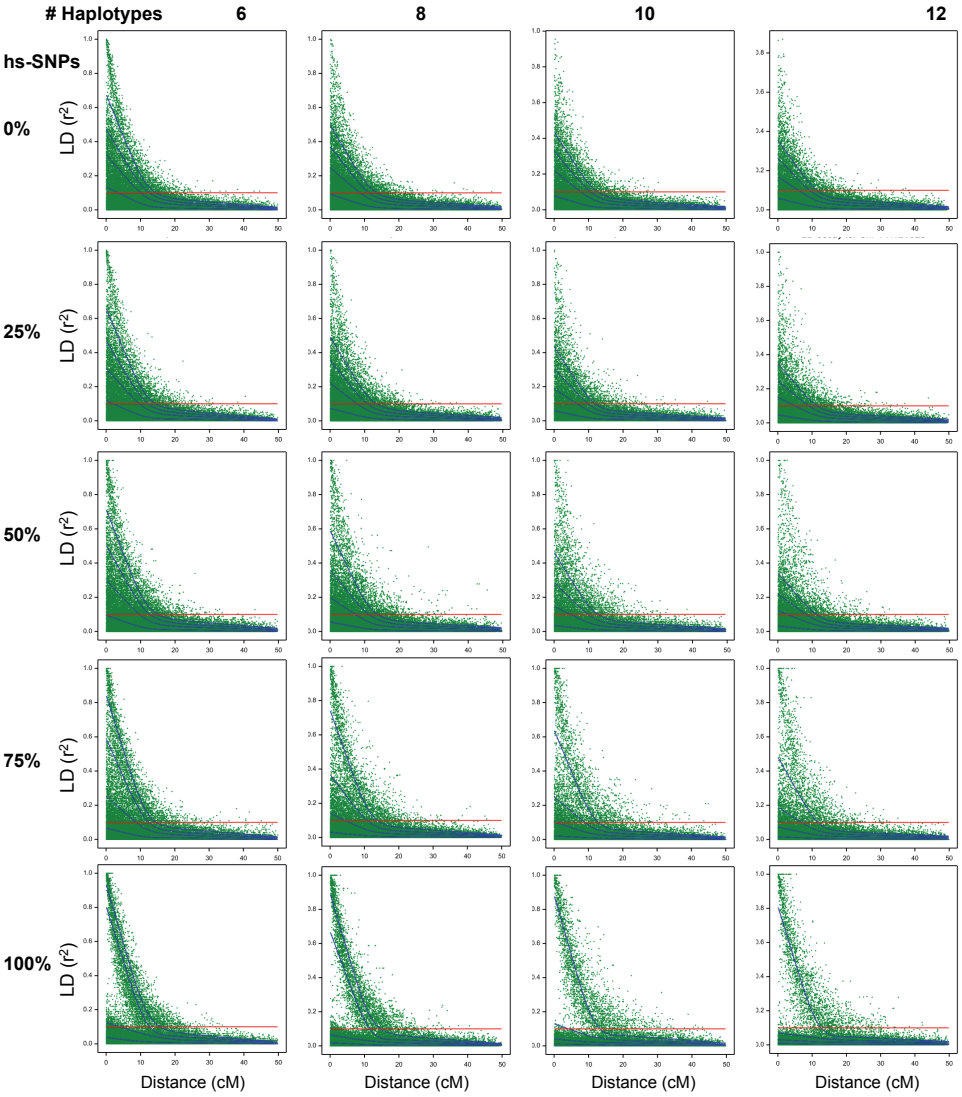

Fig. S2.

LD-decay plots of simulated data underlying the LD-decay estimates shown in Table 2. Each plot represents 1 chromosome of one of the 20 simulated datasets differing in the percentage of haplotype specific SNP and number of haplotypes. In each graphs splines are fitted on four percentile (50%, 80%, 90% & 95%).

Chapter 4

GWAS in tetraploid potato: Identification and validation of SNP markers associated with glycoalkaloid content

Peter G. Vos[1,3], M. João Paulo[2,3], Peter M. Bourke[1], Chris A. Maliepaard[1], Richard G.F. Visser[1,3], Fred A. van Eeuwijk[2], Herman J. van Eck[1,3]

[1] Plant Breeding, Wageningen University & Research , P.O. Box 386, 6700 AJ Wageningen, The Netherlands
[2] Biometris, Wageningen University & Research, P.O. Box 16, 6700 AA Wageningen, The Netherlands
[3] Centre for BioSystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands

Submitted to Theoretical and Applied Genetics
Supplementary files can be downloaded from http://dx.doi.org/10.18174/387985

Abstract

Genome wide association studies (GWAS) are a useful tool to unravel the genetic architecture of complex traits, but the results can be difficult to interpret. Population structure, genetic heterogeneity and rare alleles easily result in false positive or false negatives associations. This paper describes the analysis of a GWAS panel combined with three bi-parental mapping populations to validate GWAS results, using phenotypic data for steroidal glycoalkaloid (SGA) accumulation and the ratio between the two major glycoalkaloids αsolanine and αchaconine in tubers. SGAs are secondary metabolites in the *Solanaceae* family, functional as a defence against various pests and pathogens and in high quantities toxic for humans. With GWAS we identified five quantitative trait loci (QTL) of which *Sga1.1, Sgr8.1* and *Sga11.1* were validated, but not *Sga3.1* and *Sgr7.1*. In the bi-parental populations *Sga5.1* and *Sga7.1* were mapped, but these were not identified with GWAS. The QTLs *Sga1.1, Sga7.1, Sgr7.1, Sgr8.1* co-localize with genes *GAME9, GAME 6 / GAME 11, SGT1 and SGT2*, respectively. For other genes involved in SGA synthesis no QTLs were identified. The results of this study illustrate a number of pitfalls in GWAS of which population structure seems the most important. We also show that introgression breeding for disease resistance has introduced new haplotypes to the gene pool involved in higher SGA levels in certain pedigrees. Finally we show that high SGA levels remain unpredictable in potato but αsolanine / αchaconine ratio has a predictable outcome with specific *SGT1* and *SGT2* haplotypes.

Introduction

Genome wide association studies (GWAS) are widely used for the identification of SNP-markers associated with phenotypic variation in human, animals and plants. Application of GWAS in the model species *Arabidopsis thaliana* has resulted in greater understanding of the genetics of quantitative traits (Atwell et al. 2010). GWAS has also resulted in the identification of QTLs in crop species (Breseghello and Sorrells 2006; Cockram et al. 2010; Kraakman et al. 2004; Long et al. 2013; Malosetti et al. 2007; Riedelsheimer et al. 2012; Zhao et al. 2011). The main advantages of GWAS are that a wide genetic variation can be analysed simultaneously and that it utilizes all historical recombinations, resulting in an increased resolution for QTL detection when compared to traditional bi-parental linkage mapping. Nevertheless, GWAS suffers from several disadvantages not present in mapping populations (Korte and Farlow 2013). Firstly, population structure can result in the identification of false positive associations. Correction for relatedness between genotypes can solve this problem partly, however it may also result in not identifying true-positives. Secondly, the geometric distribution of the population allele frequency of SNP markers indicates that many rare SNP will have a reduced power in GWAS as compared to SNP allele frequencies in bi-parental populations. And finally, genetic heterogeneity, where different loci or alleles may lead to similar phenotypes, complicates the detection of individual QTLs.

Despite these disadvantages, association studies still offer the potential to dissect the genetics of complex traits, with the ultimate goal to improve breeding efficiency through the use of molecular markers. In the last decade several association studies have been conducted in potato (Baldwin et al. 2011; D'hoop et al. 2014; D'hoop et al. 2008; Li et al. 2010; Lindqvist-Kreuze et al. 2014; Malosetti et al. 2007; Rosyara et al. 2016; Schönhals et al. 2016; Urbany et al. 2011). All of these studies found significant marker-trait associations, but differ in number of markers and marker types used. For example, in the studies of D'hoop et al. (2008, 2014) a relatively high number of markers were used, however the multi-locus AFLP marker system is not easily simplified into a single locus assay for follow-up studies or breeding. Other studies (Baldwin et al. 2011; Li et al. 2010; Schönhals et al. 2016; Urbany et al. 2011) focused on a fixed number of several candidate genes, and as a consequence only a relatively small portion of the genome is taken into account. These studies can therefore technically not be considered as true genome-wide. Recently two SNP arrays have been developed for potato, a 10K SNP array, known as the SolCAP 8303 SNP array (Felcher et al. 2012) and a 20K SNP array as described by Vos et al. (2015). Both arrays offer genome-wide coverage of SNPs and should be able to capture the genetics underlying complex traits. However, two studies using the SolCAP array presented fewer significant marker trait associations than expected (Lindqvist-Kreuze et al. 2014; Rosyara et al. 2016). In this study we explore

the potential of the other array (Vos et al. 2015) using steroidal glycoalkaloid content in potato tubers as an example.

Steroidal glycoalkaloids (SGAs) are secondary metabolites and abundantly present in the *Solanaceae* family. These SGAs function (mainly in leaf tissue) as a defence against different plant pathogens such as insects (Nenaah 2011) and fungi (Hoagland 2009). In contrast to this beneficial property for the plant, the potato SGA's α-solanine and α-chaconine can have toxic effects on humans upon consumption (Friedman 2006). Therefore, a threshold of 200 mg kg^{-1} of tuber fresh weight is set as upper limit of total SGA content. In order to breed for varieties, not exceeding this legal threshold, the understanding of the genetics of glycoalkaloid accumulation is helpful. Previous mapping studies identified several QTLs for SGA accumulation in potato. While most studies measure SGA levels in potato leaves (Manrique-Carpintero et al. 2013; Manrique-Carpintero et al. 2014; Medina et al. 2002; Ronning et al. 1999; Sagredo et al. 2006; Sagredo et al. 2011; Yencho et al. 1998), a few have also been conducted on SGA accumulation in tubers (Mariot et al. 2016; Sørensen et al. 2008; Valcarcel et al. 2014). The majority of these mapping studies involved hybrids with wild species like *S. sparsipilum* (Sørensen et al. 2008) *S. berthaulthii* (Yencho et al. 1998), *S. phureja* (Medina et al. 2002), *S. commersonii* (Carputo et al. 2003) and *S. chacoense* (Manrique-Carpintero et al. 2014; Medina et al. 2002; Ronning et al. 1999; Sagredo et al. 2006; Sagredo et al. 2011). This suggests the involvement of alleles originating from wild species in the accumulation of SGA content in the contemporary potato germplasm. Six mapping studies reported a QTL for accumulation of SGAs on chromosome *1* (Hutvágner et al. 2001; Manrique-Carpintero et al. 2014; Ronning et al. 1999; Sagredo et al. 2011; Sørensen et al. 2008; Yencho et al. 1998). Additionally, α-solanine and α-chaconine QTLs have been mapped on chromosomes *6* and *11* in two studies (Manrique-Carpintero et al. 2014; Yencho et al. 1998) and on chromosomes *4*, *8* and *12* (Yencho et al. 1998){Yencho, 1998 #27;Yencho, 1998 #55}. However, the only QTL found for accumulation of α-solanine and α-chaconine in tubers is on chromosome *1* (Sørensen et al. 2008). The major genes involved in the SGA biosynthetic pathway in *Solanaceae* have recently been published (Itkin et al. 2013). In their paper six genes involved in glycoalkaloid metabolism in potato are described with four and two *GAME* (GLYCOALKALOID METABOLISM) genes located on chromosome *7* and *12* respectively. The cluster on chromosome *7* includes two of the three known SGT-genes in the potato genome. It has been shown that SGT-genes are responsible for the final glycosylation steps in the SGA pathway (McCue et al. 2011; McCue et al. 2005; Moehs et al. 1997). In the paper of Cárdenas et al. (2016) *GAME9* was postulated as a major regulator of the SGA pathway which co-localized with the QTL on chromosome *1* from Sørensen et al. (2008). From these studies the key-metabolic and regulator genes of the SGA pathway are known, however to be able to apply this knowledge to marker assisted selection in a breeding program, it is essential to know which allelic variants of these genes are responsible for variation

in SGA content in potato tubers. We therefore used a GWAS approach combined with QTL- mapping from bi-parental segregating populations to identify natural variation related to glycoalkaloid accumulation in potato tubers.

Materials and Methods

Plant materials

For the genome wide association study a variety panel of 275 tetraploid genotypes was used (**Supplementary file 1**). This is a subset of the variety panel described in Vos et al. (2015)/**Chapter 2**. In addition three bi-parental segregating populations were used, **(1)** a tetraploid bi-parental population Altus × Colomba of 87 genotypes, which is a subset from Bourke et al. (2015) (hereafter referred to as A × C), **(2)** a tetraploid bi-parental population Altus × KA 2004-4057 of 34 genotypes (hereafter referred to as A × K) and **(3)** a diploid bi-parental population SH 83-92-488 × RH 89-039-16 (Van Os et al. 2006) of 157 genotypes (hereafter referred to as SH × RH).

Phenotypic data collection

Phenotypic data for the variety panel has been collected from three sources. Field trials were performed in 2008 and 2009. In 2008 187 genotypes and in 2009 132 genotypes were phenotyped for α-solanine and α-chaconine content as described below. For another 143 varieties the phenotypic values of total glycoalkaloid content were collected either as part of VCU testing (Value for Cultivation and Use; N=84), as a requirement before adoption in the National List (Anonymous 2015) and/or were collected as multi-year, multi-location data from breeding programs as described in D'hoop et al. (2011). The tetraploid population A × C was phenotyped three times: in 2011 and two replicates on a field trial in 2012. The diploid population SH × RH and the tetraploid population A × K were both grown in field trials once, in 2009 and 2011 respectively.

TGA measurements

As a first step of TGA extraction 5 kg of tubers including the skin were ground and the liquid was separated. Subsequently 300 mg of the potato juice was put in a 15 ml plastic centrifuge tube and standard TGA extraction buffer was added (5% acetic acid supplemented with 20mM Na-1-heptanesulfonate). Samples were incubated overnight while shaking at 150 strokes/minute followed by centrifugation at 9000 xg for 10 minutes. Supernatant of the extract was filtered using a Pall Acrodisck GHP 0.45 µm syringe filter.

Analysis was carried out on an isocratic online-SPE-HPLC system. The online-SPE system contains Oasis HLB Prospect-2/Symbiosis 10x2mm cartridges (Waters, Milford Massachusetts, USA). Cartridges are replaced preventively every 96 cycles. The cartridge

was washed at a flow rate of 1 ml/min for 4 minutes with 100% acetonitrile followed by a 3 minutes wash with ultrapure water after which 850µl of TGA extract was loaded onto the cartridge. The cartridge was subsequently washed for 30 seconds with water, 2 minutes with wash buffer 1 containing 25% acetonitrile / 1.5% NH_4OH, 30 seconds with ultrapure water followed by 4 minutes with wash buffer 2 containing 15% acetonitrile / 10mM phosphate buffer pH 7.6. TGA was retrieved from the cartridge by the mobile phase of the HPLC system. Separation of α-solanine and α-chaconine was carried out on a Hypersil ODS 250 mm × 4.6 mm 5 µm C18 column with a Hypersil ODS C18 5.0 × 4.6 mm guard column (Thermo scientific, Waltham, USA) at a temperature of 34ºC and a flow rate of 1.6 ml/min. The mobile phase consisted of a 60/40 mixture of 100% acetonitrile and 10mM phosphate buffer pH 7.6 degassed using the degas function of a ultrasonic cleaner (VWR international, Leuven, Belgium) and filtered over a 0.45µm PVDF filter (Waters, Milford Massachusetts, USA). Quantification of α-solanine and α-chaconine was based on the peak absorbance area at a wavelength of 202 nm. Total SGA (equal to the sum of α-solanine and α-chaconine), α-solanine and α-chaconine content was reported in mg kg^{-1} (PPM).

Marker data

The variety panel and both tetraploid bi-parental populations (A × C and A × K) were genotyped with the SolSTW 20K Infinium array (Vos et al. 2015). In the two bi-parental populations 7150 and 7144 markers segregated respectively. In the variety panel an allele frequency cut-off of 1.25% was used, which implies for tetraploids that approximately 5% of the varieties are positive for the minor allele, mostly in simplex condition. This resulted in 11674 SNP markers that were polymorphic in the entire panel of 275 genotypes (11147 SNPs in the subset of 132 genotypes). The SNP markers were used according to the reference genome where the SNP dosage used is the dosage of the non-reference SNP-allele. The diploid mapping population was genotyped with AFLP. The genetic map of this population (Khan 2012; Van Os et al. 2006), could be used to convert the genetic positions of marker bins into physical coordinates (Sharma et al. 2013).

QTL detection

Prior to the GWAS and QTL mapping, Best Linear Unbiased Estimates (BLUE) were calculated using restricted maximum likelihood (REML) with the model as used by D'hoop et al. (2008, 2014) and shown here:

Response = year + origin + (year.origin) + genotype

An initial GWAS was performed using a subset of 132 genotypes for which repeated phenotypic measurements were available for both α-solanine and α-chaconine.

Subsequently the complete set of 275 genotypes was used, including all 187 genotypes from the field trial and the 88 additional genotypes tested for their VCU.
For the GWAS both a naive and a kinship-corrected model were applied. For the naive GWAS a regression model was used for an association between allele dosage and glycoalkaloid content. A mixed model was used to correct for relatedness among these 132 or 275 varieties, to prevent false positive detection due to population structure. A kinship matrix was calculated using ecological distance in GenStat ($1-|x_i-x_j|/r$, unless $x_i=x_j=0$, where x_i and x_j are allele dosages and r is the range). For this purpose 710 independent markers have been selected, which resulted in the identification of three subpopulations named the "Agria", "Starch" and "Rest" group as defined by D'hoop et al. (2008). A principal coordinate analysis based on these 710 markers was shown before (**Chapter 3**). For the QTL analysis in the tetraploid bi-parental populations single-marker regression using marker dosages was used, similar to the naive GWAS model.

The naive model (regression)
Trait (y) = Marker (m) + Residual (where *var ()* = $I\sigma_{\varepsilon}^{2}$

The model correcting for relatedness (mixed model)
Trait (y) = Marker (m) + genotype + residual where var (genotype) = $K\sigma_{g}^{2}$ *and K= Kinship matrix*

Underlined terms are considered as random and not underlined terms as fixed effects.
For QTL mapping in the diploid bi-parental population SH × RH MapQTL (Van Ooijen 2004) was used. Interval mapping was applied after which a major QTL was used as a co-factor in the analysis (i.e. a restricted MQM was applied).

Backward selection
The GWAS reports single marker analyses. A backward selection procedure was used to combine all significant markers in a multi-locus model. In this procedure we first removed redundant markers by picking the most significant marker in a set of markers that are in linkage disequilibrium (LD) ($r^2 > 0.9$). In the backward selection procedure the least significant marker in the model was removed until all markers were significantly ($p < 0.05$) contributing to the model.

Results

Phenotypic evaluation

Phenotypic data for this study was collected from several sources: **(1)** a designated field trial, **(2)** official VCU data published by Plantum.nl (Anonymous 2015) to allow admittance to the National list (the legally permitted threshold of SGA may not exceed 200 mg kg^{-1}) and **(3)** historical multi-year, multi-location data collected from breeding programs (D'hoop et al, 2011). Correlations of the overlap between the different data origins resulted in relatively high R^2-values (**Fig. S1**). The R^2 between VCU data and our field trial was 0.72 (n=54) and the R^2 between historical data and VCU data was 0.85 (n=13). There was no overlap between the field trial and historical breeding data. Because of these high correlations we were confident that merging the different datasets would likely increase the power for QTL detection.

Working with a variety panel on a trait for which high values are legally prohibited, it is obvious that the number of varieties with high SGA content is limited. Nevertheless, varieties exceeding the 200 mg kg^{-1} threshold were identified, which, apart from Lenape, were all varieties from the starch industry and not used for human consumption such as Festien, Allure, Astarte, Elkana, Mantra, Kuras and Mercator. Having such a low number of high SGA varieties in our variety panel results in an extremely skewed distribution as shown in **Fig. 1** and **Fig S2**. A similarly skewed distribution of SGA phenotypic values is observed in the bi-parental populations A × K, A × C and SH × RH (**Fig. S2**), although not as extreme as in the variety panel. In order to obtain a normal distribution of the trait values required for a correct analysis, a ^{10}log-transformation was applied, and to be consistent, this transformation was applied for all datasets. After a ^{10}log-transformation, BLUEs (best linear unbiased estimates) were calculated for the variety panel and the A × C population using REML. For both populations no significant origin, year or location effects were identified. The A × K and SH × RH were only phenotyped once; therefore REML could not be applied. The ^{10}log-transformation resulted in a more normal distribution of trait values suitable for GWAS (**Fig. 1**).

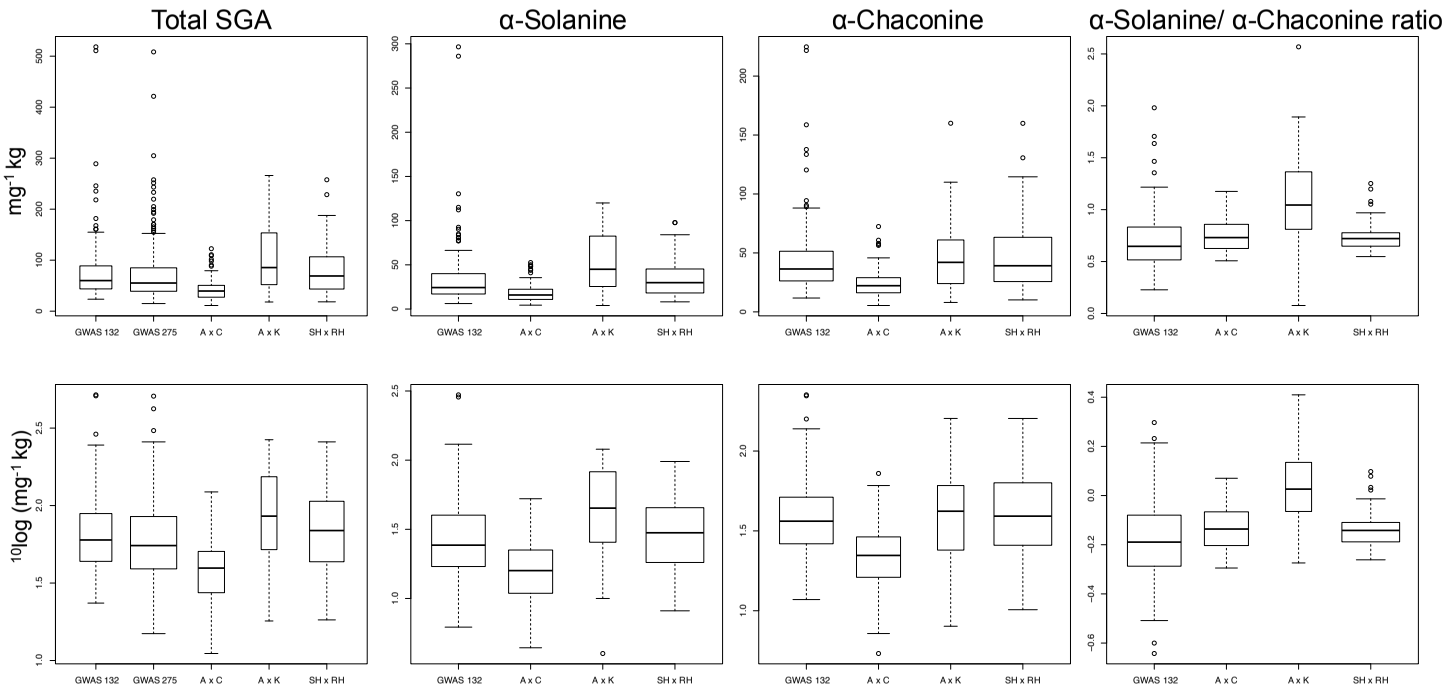

Fig. 1 Boxplots showing statistical distribution of all traits.

Both the distribution for the raw data (mg kg^{-1}) and ^{10}log-transformed data are shown. The width of the boxplot is proportional to the population size.

Table 1. Heritabilities.

Broad sense heritabilities (H^2) are shown for the glycoalkaloid traits in mg kg^{-1} (untransformed) and the ^{10}log transformation of mg kg^{-1}.

Population	# Individuals	Trait	H^2 mg kg^{-1}	H^2 ^{10}log(mg kg^{-1})
Variety panel	n = 132	α-Solanine	0.58	0.73
		α-Chaconine	0.58	0.56
		Total glycoalkaloids	0.58	0.65
		Ratio (α-Solanine/α-Chaconine)	0.26	0.53
Variety Panel	n =275	Total glycoalkaloids	0.57	0.70
A x C (4x)	n = 92	α-Solanine	0.67	0.74
		α-Chaconine	0.64	0.68
		Total glycoalkaloids	0.65	0.70
		Ratio (α-Solanine/α-Chaconine)	0.77	0.78
A x K (4x)	n = 32	All traits	na[a]	na
SH x RH (2x)	n = 157	All traits	na[a]	na

a) The populations A × K and SH × RH were phenotypes once, therefore no heritability could be calculated

The broad sense heritability (H^2) was calculated (**Table 1**)**.** Highly similar estimates (H^2 = 0.58) were observed for α-solanine, α-chaconine and total SGA content in the variety panel. In the A × C population the heritabilities for the different SGA traits ranged between 0.64 and 0.67 (**Table 1**). The ^{10}log-transformation resulted in a slight increase of the heritabilities for α-solanine and total SGA content (**Table 1**). The heritability of α-solanine/α-chaconine ratio was relatively low (H^2 = 0.26) in the variety panel, but H^2 = 0.77 was observed in the A × C population. Also for the ratio the ^{10}log-transformation resulted in an increase of the heritability in the variety panel to 0.53.

In **Fig. 1** boxplots show the statistical distribution of all traits in all populations this study for both the raw data (mg kg^{-1}) and after the ^{10}log-transformation. The maximum SGA content observed in the variety panel is 518 mg kg^{-1}. Some progeny from two bi-parental populations also exceeded 200 mg kg^{-1}, of which the SH × RH population shows a clear transgressive segregation (**Fig. S2**). The third population A × C does not show transgressive segregation and the complete population remains well below the 200 mg kg^{-1}-threshold. The ratio between α-solanine and α-chaconine in the majority of the varieties is below 1 indicating that more α-chaconine is accumulating than α-solanine. In **Table S1** the correlation between the different SGA traits for the different populations is shown. For all populations a very high correlation between α-solanine, α-chaconine and total glycoalkaloid content is observed, while the correlation between ratio and level of SGAs is less significant or even absent.

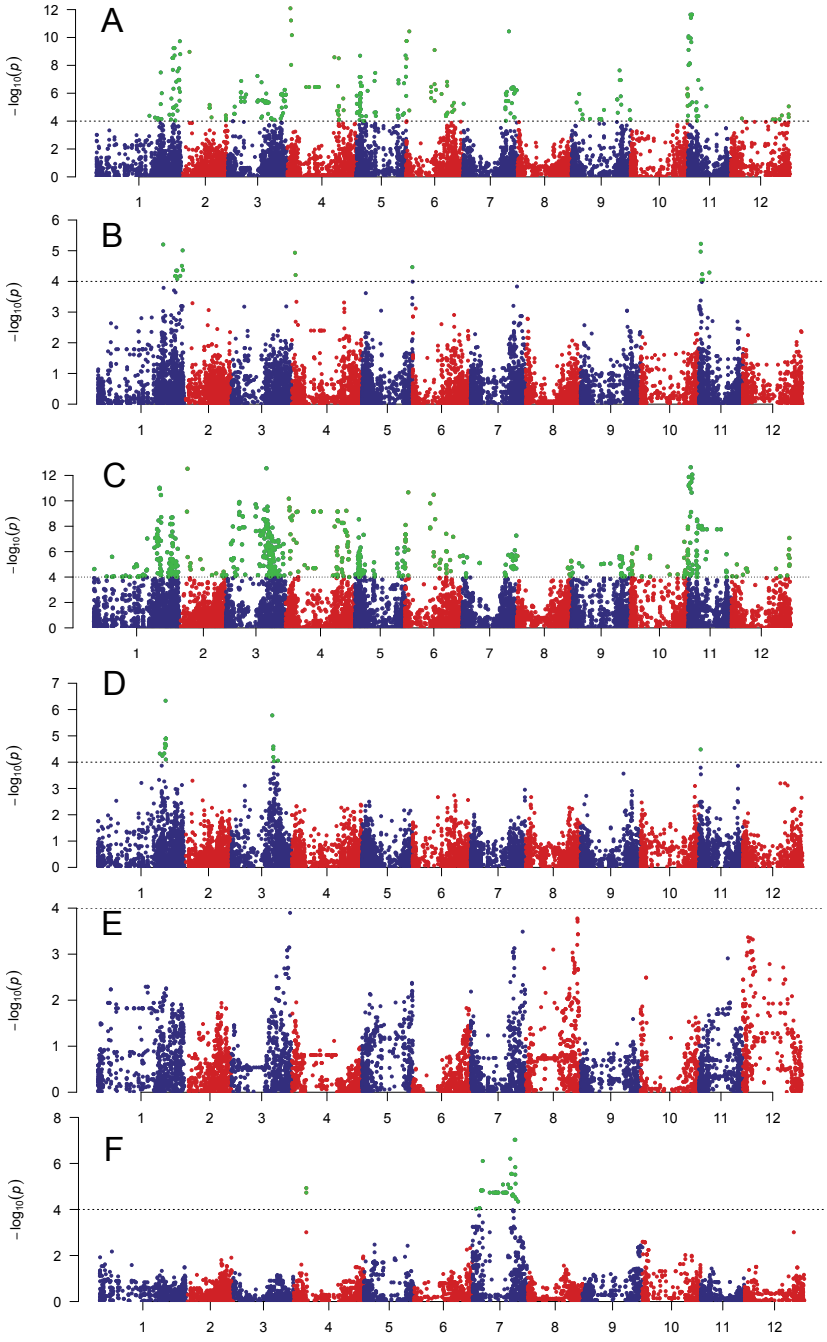

Fig. 2 Manhattan plots for total SGA content.

Manhattan plots of a (A) naive GWAS with 132 genotypes, (B) kinship corrected GWAS with 132 genotypes, (C) naïve GWAS with 275 genotypes (D) kinship corrected GWAS with 275 genotypes, (E) QTL analysis of bi-parental mapping population A × C and (F) QTL analysis of bi-parental mapping population A × K. Dotted horizontal line is at the multiple-testing threshold of $-^{10}\log(p)$ of 4. All green dots are significant according to this threshold.

QTL identification for glycoalkaloid content

An initial GWAS was performed with 132 genotypes, because these genotypes have been on the field trial in two consecutive years and both α-solanine and α-chaconine phenotypes were available. Analysis of α-solanine and α-chaconine contents (as separate traits) gave almost identical results as total SGA contents, which is due to their high correlations. Separate α-solanine and α-chaconine data are available for only the 132 varieties of the basic panel and in view of the high correlations we only show results of total SGA contents. As a second step we performed a GWAS on the complete panel of 275 varieties. **Fig. 2a** shows the results of a naive GWAS on 132 genotypes, where 367 markers exceeding the multiple testing threshold of $-^{10}\log(p) \geq 4$ (indicated with green dots). The same naive analysis with all 275 genotypes resulted in 651 significantly associated markers (**Fig 2c**). The maximum $-^{10}\log(p)$ of both analyses is around 12. Population structure strongly affects the results of GWAS, and therefore both naive and kinship corrected results of GWAS are presented in **Fig. 2a** and **b** for the GWAS with 132 genotypes and **Fig. 2c** and **d** for the GWAS with 275 genotypes. A kinship-corrected GWAS of the initial (N=132) and the total panel (N=275) both identified 21 significantly associated markers (listed in **Supplementary file 2**). The strong reduction in the number of significant marker trait associations suggests a strong correlation between trait values and population structure. Indeed, **Fig. 3a** confirms this confounding effect, as higher SGA values are overrepresented among varieties that belong to the "Starch" subpopulation. More striking is that the 21 significantly associated markers from the analysis with 132 genotypes with replicated phenotypes and the 21 SNPs from the complete set of 275 plants are completely non-overlapping. Therefore, we also tested the subset of 143 plants for which we relied on data from the VCU test and the historical data from breeding programs. The kinship corrected GWAS identified eight SNPs exceeding the threshold (listed in **Supplementary file 2**) of which two associations were unique for this supplementary panel and six SNPs overlapped with SNPs discovered in the total panel. Remarkably for several SNPs the significance of the association and the explained variance was much higher in this complementary panel as compared to the total panel. When we compared the physical coordinates of the SNPs from the different association analyses, it was obvious that the significant markers largely came from the same genomic regions on chromosomes *1* and *11*. However unique positions were found as well for the initial panel on chromosome *4* and *5* and for the complete panel on chromosome *3*.

An overview of the SNPs and QTL regions identified in this study is presented in **Table 2**, and can be summarized as the detection of five genomic regions containing QTLs involved in amount of steroidal glycoalkaloids: *Sga1.1*, *Sga3.1*, *Sga5.1*, *Sga7.1* and *Sga11.1*.

Table 2:

Summary of the QTL and peak markers underlying the QTLs as shown in Figure 2 and 6

QTL Locus	Region (Mb)	peak SNP [a]	GWAS Panel [a]		Bi-parental Validation Populations								Candidate gene [b]
					A × C		A × K		SH × RH				
			$^{10}\log(p)$ (R^2)	Effect	$^{10}\log(p)$ (R^2)	Effect	$^{10}\log(p)$ (R^2)	Effect	Peak (cM)	LOD	Effect	
Sga1.1	chr01:63.8-87.5	PotVar0043608 (2D)	6.3 (9.3)	0.35					81.5	27.8 (56.7)	0.37	GAME 9 (25989)
Sga3.1	chr03:42.2-48.0	PotVar0068174 (2D)	5.8 (17.5)	0.12								Unknown
Sga5.1									18.2	5.9 (7.1)	-0.19	Unknown
Sga7.1	chr07:7.1-45.1	PotVar0092875 (2E)					8.1 (63.9)	0.37	38.0	6.2 (7.5)	-0.14	GAME 6 (11750) and GAME 11 (11751)
Sgr7.1	chr07:9.1-42.1	PotVar0069919 (6B)	8.5 (24.4)	-0.11								SGT 1 (11749)
Sgr8.1	chr08:48.2-50.4 chr08:47.3-54.8	PotVar0063333 (6B) PotVar0063060 (6C)	7.1 (19.7)	-0.07	9.5 (35.5)	0.21						SGT 2 (17508)
Sga11.1	chr11:2.0-11.0	PotVar0066293 (2B)	5.2 (14.5)	0.16					37.9	5.6 (15.4)	0.14	Unknown

[a]) The peak SNP underlying the QTL region, as detected in the basic or complete panel, refers to the most significantly associated SNP. The information within brackets refers to the respective Manhatten plots and GWAS panel in Fig. 2 and 6.

[b]) PGSC0003DMG numbers are added between brackets

Within the *Sga1.1* region the most significant SNPs identified in the kinship corrected GWAS across the different panels are located on chromosome *1* within superscaffold PGSC0003DMB000000095 explaining up to 21% of the variation. The 23.8Mb large region (chr01:63764681…87544718) has two sub-regions where SNPs pile up predominantly within a 4Mb region (chr01:65692910…69665033) identified in the analyses with 132 and 275 genotypes and another 7.7 Mb interval more south (chr01:80467982…87185358) only identified in the subset of 132 genotypes. Both are independent, because the markers of the sub-regions do not correlate (data not shown). Validation of *Sga1.1* with bi-parental mapping populations was not feasible with the A × C or A × K populations, because none of the significant SNPs segregated in A × C and only three segregate in A × K, but were not significant. The QTL *Sga1.1* was well confirmed in mapping population SH × RH, identifying a major QTL on SH01 explaining 56.7% (LOD= 27.8) of the variation (**Fig. 4**). Using the study of Sharma et al. (2013) the peak marker maps around the exact same region of PGSC0003DMB000000095.

The locus *Sga3.1* maps to a sharp QTL peak encompassing SNPs from a 1.7Mb interval (chr03:42232155…43921814) where PotVar0068174 is the most significantly associated SNP ($-^{10}\log(p)$ = 5.8) with SGA in the total panel, explaining 17.5% of the variation. In the subset of 132 genotypes the associations did not reach the significance threshold of 4. This QTL was not validated in any of the bi-parental mapping populations, although PotVar0068174 was segregating in both A × C and A × K.

GWAS identified PotVar0034580 as a single, significantly associated SNP ($^{10}\log(p)$ = 4.5) located on a distal position (51.70 Mb) of the south arm of chromosome *5*. We did not assign a QTL name to this putatively spurious association, because PotVar0034580 was only significant in the basic panel and not associated with SGA content in the total panel, nor in any of the mapping populations. On the north arm of chromosome *5,* however the validation population SH × RH displayed a significant QTL ($^{10}\log(p)$ = 5.9) at 18.2 cM, explaining 7.1% of the variance. The AFLP markers in the SH ´ RH population, associated with the QTL called *Sga5.1,* correspond to a position on superscaffold PGSC0003DMB000000192.

The validation population A × C population did not result in any significant QTL (**Fig. 2e**), while the validation population A ´ K displayed a highly significant QTL, called *Sga7.1*, on chromosome *7* with a $-^{10}\log(p)$ of 8.1, and explaining 63.9% of the variance. **Fig. 2f** shows the location of *Sga7.1* in a physically large interval with many SNPs. The long range LD in this interval is caused by suppression of recombination in the pericentromeric heterochromatin. The peak position consists of two co-segregating SNPs (PotVar0092875 and PotVar0115020). GWAS could not identify associated SNPs on chromosome *7* with SGA contents. This QTL was validated in the SH ´ RH population **(Fig. 4)** explaining 7.5% of the variation with a LOD of 6.2, although it maps on superscaffold PGSC0003DMB000000233 slightly more towards the chromosome end

(Sharma et al. 2013). The QTL interval of *Sga7.1* includes candidate genes *GAME6* or *GAME11*.

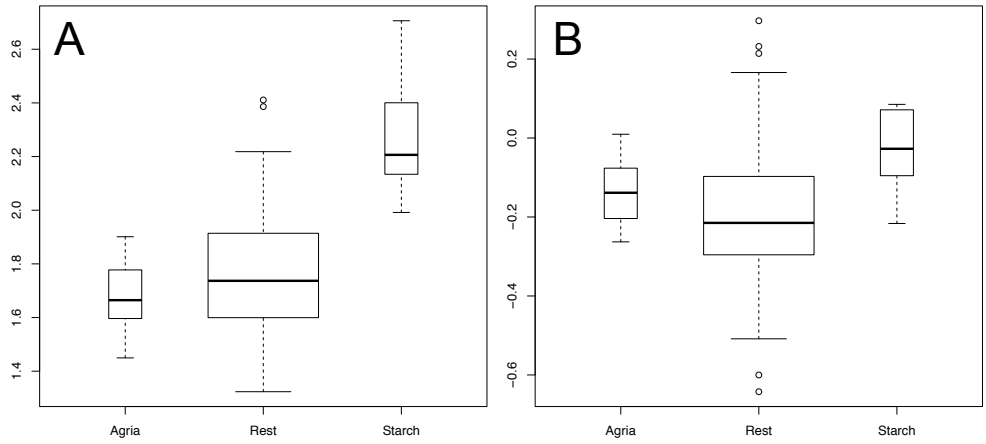

Fig. 3 Confounding of phenotypes with population structure.

Distribution of total SGA content (A) and α-solanine/α-chaconine ratio (B) over the three different subpopulations. The width of the boxplot is proportional to the population size.

On chromosome *11* a significant QTL was discovered with the subset of 132 genotypes and the complete panel (**Fig. 2b and 2d**). This QTL named *Sga11.1* maps to a 2.3Mb region (Chr11:2037454…4347636) where peak marker PotVar0066293 explains 14.5% of the variance. *Sga11.1* could be validated in the SH × RH population, where a QTL with a LOD-value of 5.6 explaining 15.4% of variation, could be mapped genetically to a position corresponding closely to superscaffold PGSC0003DMB000000133 (Sharma et al. 2013).

A backward selection procedure was used to compose a multi-locus model of non-redundant markers. The markers included in this model are listed in **Table S2** and the prediction by this model is illustrated in **Fig. 5a**. Collectively these SNPs explain 32% of the total variation. In view of the high broad sense heritability for SGA (H^2 is ranging between 0.56 and 0.73), the results of our model suggest a reasonable amount of missing heritability.

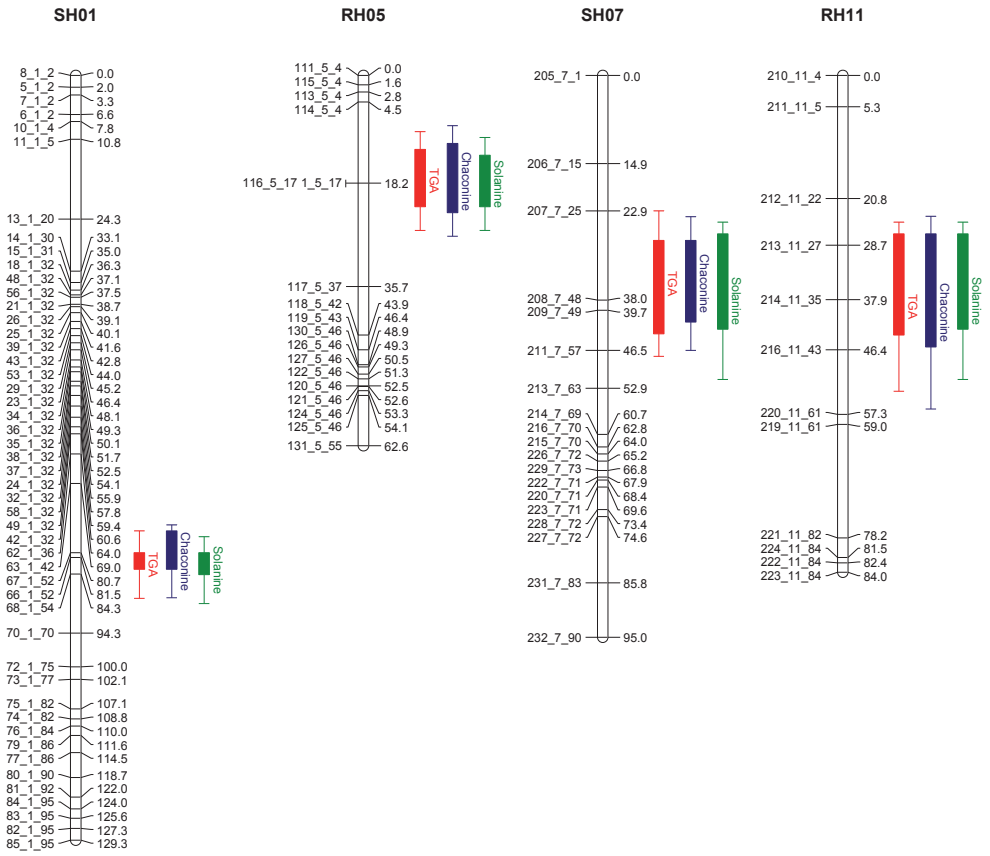

Fig. 4 QTL Mapping results of the diploid SH × RH population.

Maps of chromosomes 1, 5, 7 and 11. Marker names on the left and map positions (cM) on the right. Bars show the 2-LOD interval for all QTLs.

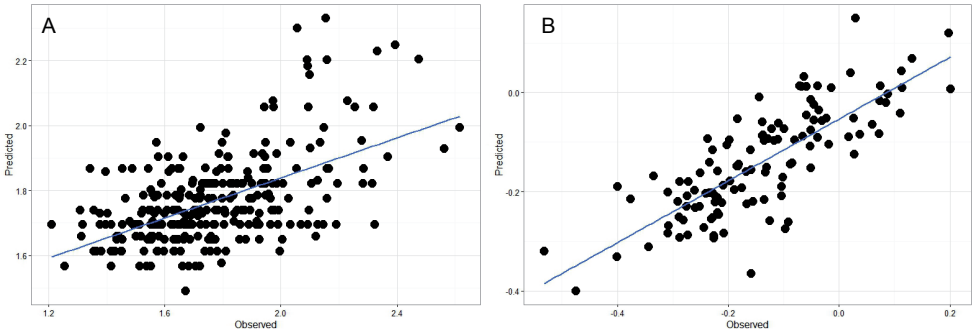

Fig. 5 Backward selection models.

Backward selection models of (A) total SGA content and (B) *α*-solanine/ α-chaconine ratio. The observed phenotypes (x-axis) are plotted against the predicted phenotypes based on a set of 5 and 6 significant SNP markers for total SGA and *α*-solanine/ α-chaconine ratio respectively.

QTL analysis for α-solanine/α-chaconine ratio

The ratio between α-solanine and α-chaconine was studied in the panel of 132 varieties only because in the additional 143 genotypes α-solanine and α-chaconine were not measured separately. In **Fig. 6** the Manhattan plots of a naive and a kinship corrected GWAS are shown. In contrast to the GWAS with total SGA, kinship correction did not cause strong reduction in the number of associated SNPs. This suggests that trait values for SGA ratio are negligibly confounded with population structure (**Fig. 3b**). A highly significant QTL named *Sgr7.1* (**Table 2**) was identified on chromosome *7* (**Fig.**

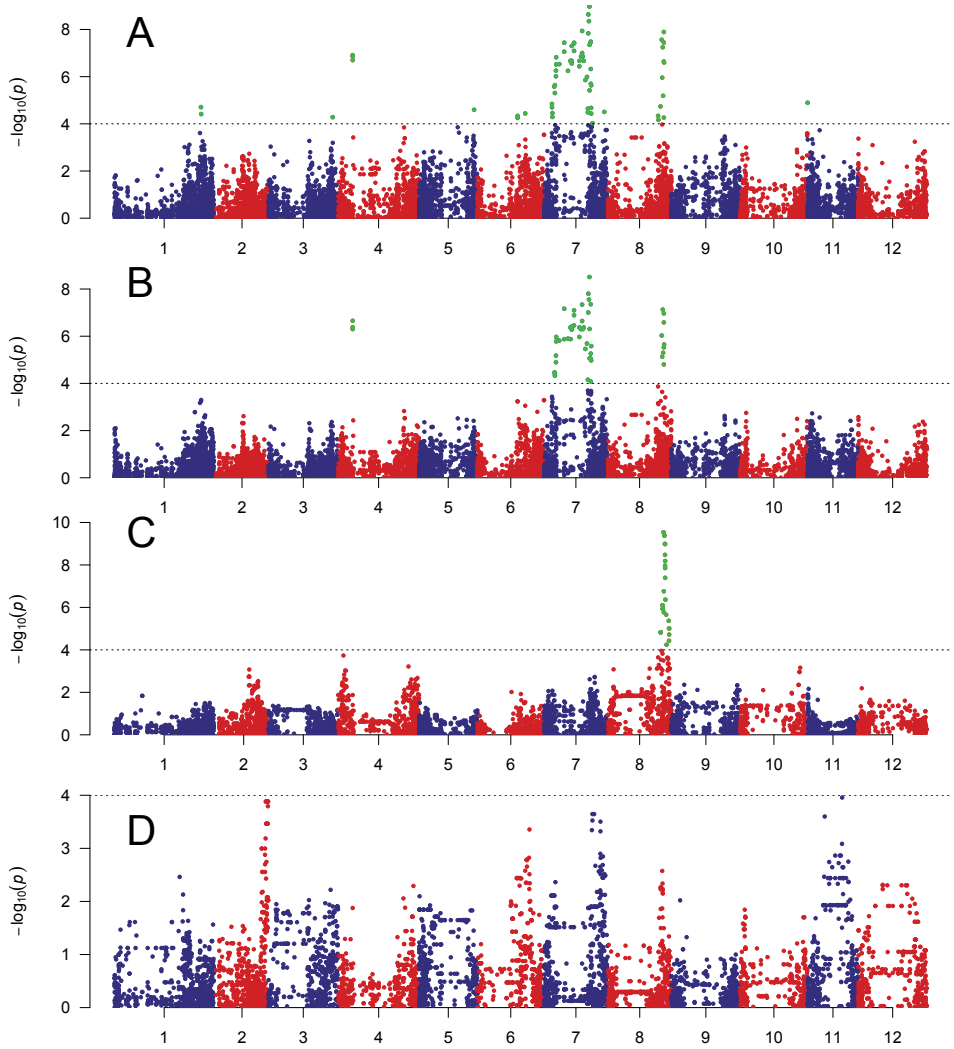

Fig. 6 GWAS/QTL-mapping results of α-solanine/ α-chaconine ratio.

Manhattan plots are shown for a (A) naïve GWAS with 132 genotypes, (B) a kinship corrected GWAS with 132 genotypes, (C) the QTL analysis in bi-parental A × C population and (D) the QTL analysis bi-parental population A × K.

6a, b). The most significantly associated SNP is PotVar0069919 ($^{10}\log(p)$ = 8.5) and explains up to 24.4% of the variation. The QTL covers a 36 Mb region, comprising approximately half of the physical chromosome (PGSC0003DMB000000684 to PGSC0003DMB000000251) and spanning the pericentromeric heterochromatin. PotVar0069919 is located at 1.5Mb distance from the most obvious candidate gene SGT1 (PGSC0003DMG400011749). Three strongly associated SNPs, shown in **Fig. 6a**, **b**, suggest the presence of a QTL on chromosome *4*. However, such isolated dots are suspicious, because they completely co-segregate with markers on chromosome *7*, and indeed a recent revision of scaffold DMB67 (Endelman and Jansky 2016) would replace these markers to chr07:26125388…26309187, right within the QTL on chromosome *7*. Many of the highly significant SNPs are present in old varieties Jaune d'Or, Yam and Myatt's Ashleaf, suggesting that this haplotype is not a recent introgression from wild species. The validation of this QTL failed because no QTLs could be discovered for α-solanine/α-chaconine ratio in the three bi-parental mapping populations. However the majority of the SNPs underlying the QTLs *Sgr7.1* (**Fig. 6a**) and *Sga7.1* (**Fig. 2f**) are the same.

The second QTL for α-solanine / α-chaconine ratio was detected on chromosome *8*. This locus, named *Sgr8.1*, is most significantly detected with PotVar0063333 ($^{10}\log(p)$ = 7.1) and explains 19.7% of the variation (**Table 2**). The SNPs significantly associated with QTL *Sgr8.1* cover a 2.1Mb region comprising the candidate gene *SGT2* (PGSC0003DMG400017508). Validation of this QTL region was possible using the A ´ C mapping population, where highly significant SNPs were observed in a much wider 7.5Mb genomic region. However the SNPs underlying this QTL are completely different in both populations (**Supplementary file 2**). In the bi-parental population a haplotype segregates, tagged by the peak marker PotVar0063060 ($^{10}\log(p)$ = 9.5), **Fig. 6c**), explains 35.5% of the variance. Remarkably, in the variety panel is PotVar0063060 a rare SNP with a population allele frequency of 0.4%, located within the open reading frame of *SGT2*. Such rare SNPs do not have the power to detect a QTL in an association panel. Graphical genotyping as performed in **Chapter 5** on the data of Uitdewilligen et al. (2013) revealed a unique rare haplotype of at least 5.8 Mb (chr08:47377386.. 53138672) (**Fig. S3**) present in the varieties Festien, Kartel and Aveka. These SNPs are identical by descent and located on an introgressed haplotype from *S. vernei*. Kartel is the male parent of variety Altus, which is the female parent of the A × C population, and therefore, the QTL could be discovered with SNPs specific to this rare haplotype. **Fig. S3** shows additional SNPs that are within the open reading frame of *SGT2* that were not implemented on the 20K array. These include PotVar0063041, a synonymous C/T SNP at coordinate 49813826, PotVar0063070, a non-synonymous G/T SNP at coordinate 49813558 and PotVar0063092, a G/A SNP causing a premature stop codon at coordinate 49813385. In the other two bi-parental populations (SH × RH

and A × K) we did not detect any significant QTLs for the ratio between α-solanine and α-chaconine.
In **Fig. 5b** the phenotypic predictions from a backward selection model is shown. This model combines a set of non-redundant significant SNPs listed in **Table S2**. The R^2 of this figure is 60 % indicating that the QTLs *Sgr7.1* and *Sgr8.1* can explain a major part of the heritability within the variety panel.

Discussion

Genome wide association studies have become a widely accepted method to explore the genetic structure of complex traits in plants. Extensive research in plants, predominantly in *Arabidopsis thaliana* have highlighted some of its advantages and limitations, nicely reviewed by Korte and Farlow (2013). They state that Arabidopsis is almost an ideal organism to conduct GWAS, because the continued self-fertilization allows repeatedly phenotyping of genetically identical individuals. From that perspective, vegetatively propagated crops such as potato should be equally suitable for GWAS. As a textbook example for GWAS we made use of publically available phenotypic data from VCU testing, and multi-year multi-location data from breeding programs in addition to a designated trial. With genotypic data collected with the 20K SolSTW array (Vos et al. 2015) we could explore the possibilities and limitations of genome wide association studies in tetraploid potato.

Population structure
In general population structure is a major obstacle in GWAS. Without a proper correction for the relationship between individuals, many spurious associations might be identified. In this study we show that total SGA content is confounded with population structure (**Fig. 3a**), resulting in many spurious associations (**Fig 2a, c**). In contrast, α-solanine/α-chaconine ratio is not confounded with population structure and the naive and kinship corrected results are almost identical (**Fig. 6**). This difference is the obvious result of selection, where selection against SGA is less important for varieties bred for starch industry, and for α-solanine/α-chaconine ratio selection is absent.
In natural populations not only selection, but also reproductive isolation due to geographic origin is involved in shaping structured populations (Kooke et al. 2016). In crop species, the various strategies of germplasm collection and controlled crosses will reduce reproductive isolation and create less structured populations. In potato no evidence has been reported so far that geographic origin contributed to population structure. Clear evidence for population structure caused by breeding towards market segments has been documented in literature (D'hoop et al. 2010; Hirsch et al. 2013) and was also described in this thesis, in **Chapter 3**. Potatoes for processing or starch

industry and fresh consumption have different requirements for trait values and therefore traits are expected to be confounded with a selection induced population structure. An approach to avoid the burden of correction for population structure was demonstrated by Zhao et al. (2011), as they performed GWAS both across and within subpopulations. Unfortunately, the subgroups in this study ("Agria" (n=28) and "starch" (n=41) are too small to perform a genome wide association study.

Missing heritability and genetic heterogeneity

When we compare the broad sense heritabilities with the variance collectively explained by all SNP markers that were included in the backward selection models, an obvious difference is found for SGA and α-solanine/α-chaconine ratio. For SGA we obtained H^2 = 65-70% while R^2=32% (**Fig. 5**), and for ratio we obtained H^2=53% while R^2=60% (**Fig. 5**). Although comparison of H^2 with R^2 is not straightforward, it is clear that the proportion of missing heritability is larger for the structure-confounded trait of SGA. This suggests that the correction for population structure has dismissed true positive markers for the structure-confounded trait SGA, and that most likely the majority of QTLs involved in α-solanine/α-chaconine ratio have been identified.

Trait values for SGA were collected from field trials, historical breeder's records and VCU documents, but these different phenotypic records relate to different variety panels. GWAS analysis of the sub-panels with either field data or the VCU data, as well as the merger of these panels into a larger GWAS panel, allowed the reproducible identification of QTL positions for SGA and α-solanine/α-chaconine ratio, as these loci were consistent across sub-panels and the total variety panel. So far the discovery of the map positions of the QTLs may seem straightforward, but at these QTL loci we lack information about the haplotype structure, or how many of the various alleles have a positive or negative influence on the trait values. This problem is illustrated by the striking observation that the 21 significantly associated SNPs detected with small sub-panel (n=132) did not overlap with the 21 significant SNPs detected with the total panel (n=275). SNPs not detected in the smaller set and become significant in the larger set or SNPs are significant in the smaller set and are no longer significant in the larger dataset. The former may be indicative for some missing heritability due to SNPs with a low allele frequency; the latter indicates that similar levels of total SGA content can be explained by different combination of alleles, i.e. genetic heterogeneity.

This study provides two examples of SNPs that were excluded from GWAS because their allele frequency was below the pre-set threshold of 1.25%. The allele frequency of PotVar0043608 in the subset of 132 genotypes is only 0.9%, but with an allele frequency of 1.6% in the supplementary panel and 1.3% in the total panel, highly significant associations with SGA contents allowed the identification of *Sga1.1*. The second example relates to the detection *Sgr8.1* with SNP PotVar0063060 in in the A × C population, associated with the haplotype derived from the variety Kartel, having a

population allele frequency of below the threshold in the GWAS panels and therefore excluded from analysis.
The genetic heterogeneity in this particular situation can be explained as the result of the different varieties found in the 132-set and the complete panel. The supplementary panel with phenotypic data from VCU records represents more modern varieties. For example two-thirds of the 132 varieties were released before 1991, whereas two thirds of N=143 supplementary panel is comprised of varieties released after 1995. Presumably, the incidence of alleles contributing to high SGA contents derived from heirlooms, has declined over the years, but introgression breeding for disease resistance have passed new alleles to the contemporary gene pool.

Candidate genes underlying QTLs for SGA content and α-solanine/α-chaconine ratio
Much progress has been made in recent years to elucidate the genes involved in the regulation and synthesis of SGAs (Cárdenas et al. 2015; Cárdenas et al. 2016; Itkin et al. 2013). This biochemical information is valuable, but it does not provide information on the loci that cause genetic differences in the amounts and composition of SGA in potato varieties. The development of marker assisted selection strategies requires the identification of relevant alleles at the QTLs involved in trait variation rather than the genes themselves.
The most significant SNP underlying *Sga1.1* is PotVar0043608 (**Fig. 2d**). This SNP is located at 160kb distance from *GAME9* (Cárdenas et al. 2016) suggesting linkage disequilibrium between a SNP and a *GAME9* allele increasing SGA contents. The QTL *Sga1.1* was validated in the SH × RH population, and also co-localizes with the QTL identified by Sørensen et al. (2008). In our study *Sga1.1* is the QTL with the largest effect, which confirms the characterisation of *GAME9* as a key regulator of the SGA pathway (Cárdenas et al. 2016). This study shows that *GAME6/GAME11* co-localize with the QTL *Sga7.1*, however the position of these genes might be confusing because the reference genome (PGSC 2011) reports a location 10Mb differing from the location reported by Itkin et al. (2013). Furthermore the QTLs *Sgr7.1* and *Sgr8.1* clearly match with the responsible genes *SGT1* and *SGT2.* Earlier reports (Krits et al. 2007) already documented on the correlation between *SGT1* and *SGT2* transcript ratio and the α-solanine to α-chaconine ratio in potato tubers. As discussed before, the validation of a *Sgr8.1* seems straightforward, nevertheless the SNPs were not validated. In the A × C validation population other SNPs were associated with α-solanine and α-chaconine ratio. This provides evidence that multiple alleles of *SGT2* are involved in modulation of the α-solanine and α-chaconine ratio in potato tubers.
Three more QTLs *Sga3.1*, *Sga5.1* and *Sga11.1* were identified at genomic positions, where no obvious candidate genes have been identified. Although *Sga3.1* was only identified by GWAS and *Sga5.1* only in one mapping population SH ´ RH, the QTL *Sga11.1* was discovered by GWAS and validated in the SH × RH population, and co-

localizes with QTLs identified in an earlier study (Manrique-Carpintero et al. 2013). On the other hand the associated SolCAP SNPs identified by Manrique-Carpintero et al. (2013) could not be reproduced in our study.

Fig. 2b suggests that PotVar0076636 and PotVar0107030 distally located on the short arm of chromosome *4* and PotVar0034580 on chromosome *5* have been ignored. These QTL regions could not be validated in the large set of 275, nor in any of the bi-parental populations. Additionally these were close to the significance threshold and were therefore assumed to be false positives.

For many of the other GAME genes postulated in literature there was no QTL identified, such as *GAME4* and *GAME12* on chromosome *12* (Cárdenas et al. 2015; Itkin et al. 2013), SSR2 on chromosome *2* (Sawai et al. 2014) and HMG1 (chromosome *2*), HMG2 (chromosome *2*) and SQE (chromosome *4*) (Ginzberg et al. 2012; Manrique-Carpintero et al. 2013; Manrique-Carpintero et al. 2014). This suggests that these genes do not display functional variation, and may be highly conserved.

Are introgression segments involved in the SGA biosynthesis?

As described in the introduction, many studies have been performed on SGA content use several wild relatives. In the paper of Vos et al. (2015), it was demonstrated that specific SNPs can be used to identify introgression segments from wild species, which were used as donor of disease resistance genes. In this study PotVar0043608 is the most significant SNP associated with *Sga1.1* on chromosome *1*. Analysis of the year of market introduction of potato varieties polymorphic for PotVar0043608 (Vos et al. 2015) has indicated that polymorphism at this SNP locus was first observed in variety Lenape (**Supplementary file 2**). Lenape is a variety with *Solanum chacoense* in its pedigree, which was withdrawn from the national variety lists because of high its SGA content (Zitnak and Johnston 1970). This study provides evidence that this specific SNP is a *Solanum chacoense* specific DNA variant. In descendants from Lenape the SNP is indicative of an introgressed haploblock responsible for an elevated SGA content.

Another example is the most significant SNP underlying the QTL involved in SGA ratio in A × C, which was first observed in the variety Kartel. Kartel has several *S. vernei* derived progenitors in the pedigree (Van Berloo et al. 2007). There are indications for yet another *S. vernei* derived haplotype underlying the QTL *Sga7.1* (but not *Sgr7.1*), because several of the highly significant SNPs in **Fig. 2a, c** and **f** originate from VTN 62-33-3, which is also a progenitor clone with several *S. vernei* derived progenitors in its pedigree.

Breeding for disease resistance has contributed new haplotypes to the gene pool. Wild species derived alleles for genes involved in SGA located on these introgressed haplotypes may increase the risk of genetic complementation. In particular the efforts to develop potato cyst nematode resistant starch varieties resulted in elevated levels of SGA. This study confirms earlier reports (Hellenäs et al. 1995; Yencho et al. 1998) proposing that

combinations of *S. tuberosum* and non-*tuberosum* alleles may be are responsible for high SGA levels in potato tubers.
In this study we illustrate the utility of GWAS in tetraploid potato to gain insight into the genetics of a complex trait. However, not all traits are equally suitable and some traits will suffer more than others from the disadvantages of GWAS. The combination of GWAS and bi-parental mapping populations seems essential to avoid incorrect interpretation of the data.

Author contribution

Conceived and designed the experiments: PGV, HJvE, MJP. Performed the experiments: PGV. Contributed data: CM and PB. Analysed the data: PGV. Wrote the manuscript: PGV, HJvE. Edited the manuscript: PMB, HJvE, FAvE, RGFV.

Acknowledgments

PGV is supported by a grant of CBSG (Centre for BioSystems Genomics) and by potato breeding companies Agrico Research B.V., Averis Seeds B.V., HZPC Holland B.V., KWS POTATO B.V. and Meijer B.V. We specially thank Averis seeds B.V. for providing phenotypes from the A × C population. Genetic data from the A × C population was part of the TKI polyploids project "A genetic analysis pipeline for polyploid crops" (project number BO-26.03-002-001)

Table S1.

Correlations between SGA-traits. Correlations (r) between α-solanine, α-chaconine, total glycoalkaloid content and ratio (α-solanine/α-chaconine) are shown

		α -solanine	α -chaconine	TGA
Cultivar Panel (n = 132)	α -chaconine	0.88**		
	TGA	0.97**	0.96**	
	Ratio	0.35**	-0.02	0.19*
A × C	α -chaconine	0.92**		
	TGA	0.98**	0.98**	
	Ratio	0.44**	0.09	0.26**
A × K	α -chaconine	0.81**		
	TGA	0.95**	0.95**	
	Ratio	0.18	-0.33	-0.09
SH × RH	α -chaconine	0.96**		
	TGA	0.99**	0.99**	
	Ratio	0.05	-0.19*	-0.09

*p-value<0.01, ** p-value < 0.001

Table S2

SNP markers included in the backward selection models

trait	mkid	logp	explv	effect	totexplv
SGA	PotVar0043360	1,36	12,07	0,10	31,4
SGA	PotVar0043608	2,06	9,34	0,19	31,4
SGA	PotVar0066209	4,65	2,47	0,13	31,4
SGA	PotVar0068174	3,34	17,55	0,06	31,4
SGA	solcap_snp_c2_55072	2,74	9,69	0,04	31,4

trait	mkid	logp	explv	effect	totexplv
ratio	PotVar0069919	3,80	23,37	-0,06	59,6
ratio	PotVar0092712	2,01	11,33	-0,02	59,6
ratio	PotVar0096222	1,41	10,41	-0,02	59,6
ratio	PotVar0102788	4,33	13,35	0,04	59,6
ratio	solcap_snp_c2_15925	1,50	10,34	-0,06	59,6
ratio	solcap_snp_c2_34710	3,26	17,19	-0,04	59,6

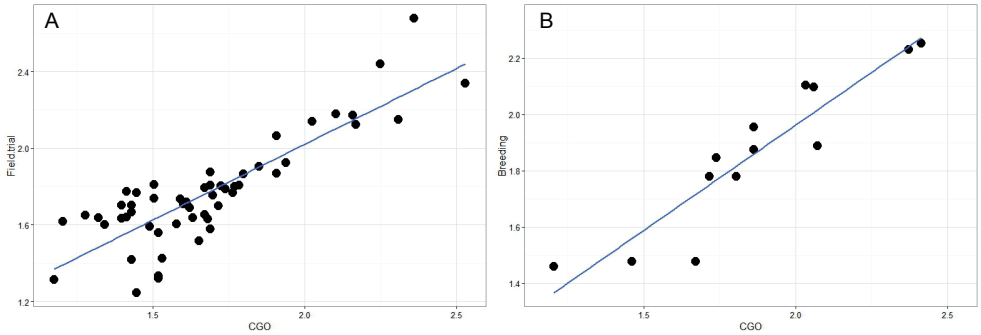

Fig. S1

Correlation of overlapping genotypes in (A) VCU/Field trial and in (B) VCU/Multi-Year-Multi-Location

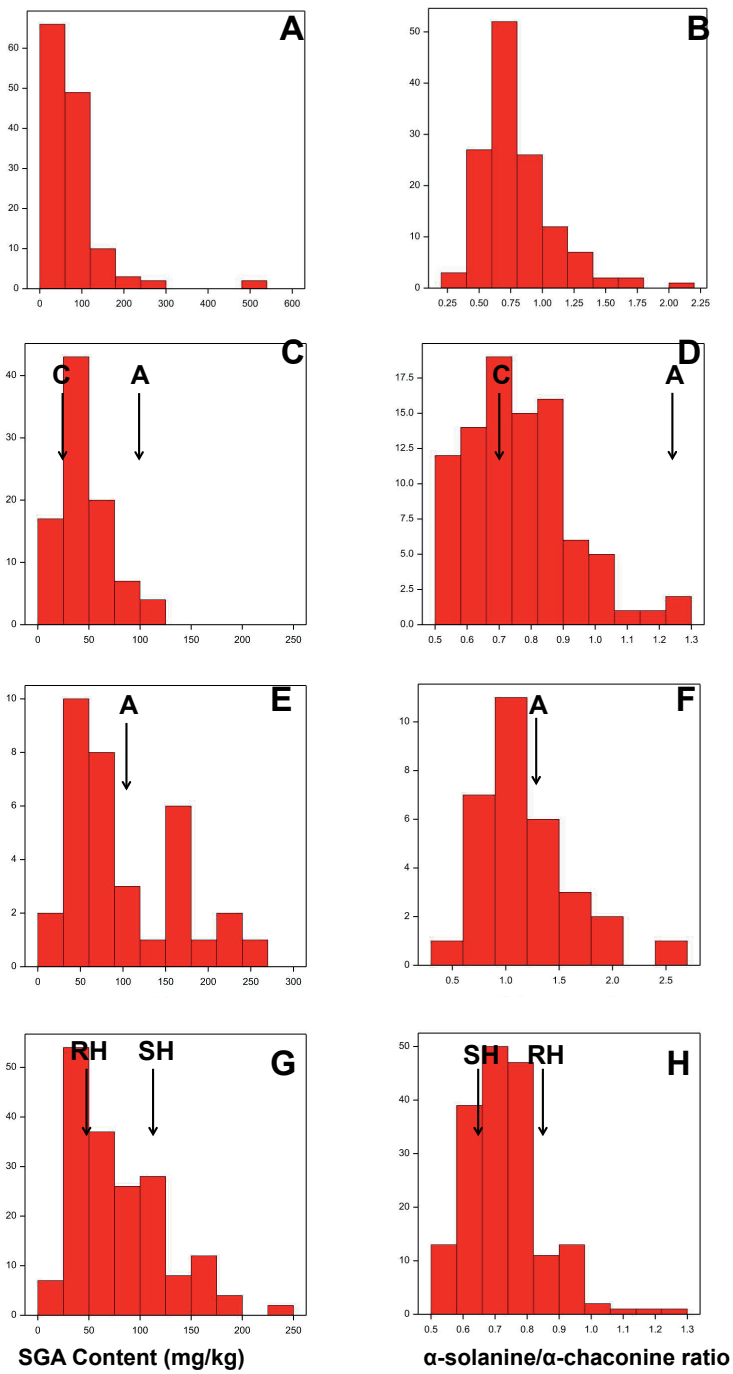

Fig. S2

Phenotypic distributions of four populations, (A & B) Variety panel, (C & D) A x C, (E & F) A x K and (G & H) SH x RH. The left panels (A, C, E, G) show total glycoalkaloid content; the right panels (B, D, F, H) show the ratio between α-solanine and α-chaconine

linkage group	Potvar	superscaffold	Position	Monoploid MV5 (1-3 551)	1256A(23) = BLACK 1256	Ackersegen	ADRETTA	Agria	Ajiba	Albion	Alpha	Amyla	Anosta	Arran Chief	Arran Pilot	Arrow	Aurora	Avenance	Ballydoon	Belle de Fontenay	Bintje	Biogold	Charlotte	Cherie	Civa	Clivia	Daisy	Ditta	EHUD	Eos	Estima	Exquisa	Felsina	Festien	Fianna	Fontane	Frieslander	Gladstone	Golden Wonder	Great Scot	Hansa	Herald	Hermes	HINDENBURG	Home Guard	Industrie	Innovator	Irish Queen	Kartel	Katahdin	Kepplestone Kidney	Kerpondy	KURAS	Laura	LIBERTAS	Markies	Mercator	Mondial	MPI 19268	Nicola	Nomade	Obelix	Pentland Dell	Picasso	Samba	Shamrock	Tasso	Tinwald's Perfection	Toyoshiro	Ultimus	Umatilla Russet	USDA 96-56	VE 70-9	VE 71-105	VE 74-45	Victoria	Vivaldi	VK 69-491	Voran	VTN 62-33-3	Winston	Wisent	Y 66-13-636	Yam	VITELOTTE NOIR	Princess	Aveka
chr08	PotVar0103355	PGSC0003DMB000000303	47377386	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0103406	PGSC0003DMB000000303	47377830	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0103435	PGSC0003DMB000000303	47378158	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	1
chr08	PotVar0063160	PGSC0003DMB000000144	49176003	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063280	PGSC0003DMB000000144	49173569	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063289	PGSC0003DMB000000144	49173444	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063291	PGSC0003DMB000000144	49173402	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063293	PGSC0003DMB000000144	49173398	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063307	PGSC0003DMB000000144	49173214	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063329	PGSC0003DMB000000144	49172889	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063331	PGSC0003DMB000000144	49172802	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063334	PGSC0003DMB000000144	49172669	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063337	PGSC0003DMB000000144	49172652	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063341	PGSC0003DMB000000144	49172556	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063342	PGSC0003DMB000000144	49172526	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063041	PGSC0003DMB000000144	49813826	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063060	PGSC0003DMB000000144	49813646	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063070	PGSC0003DMB000000144	49813558	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063092	PGSC0003DMB000000144	49813385	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063369	PGSC0003DMB000000144	48939745	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063371	PGSC0003DMB000000144	48939715	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063381	PGSC0003DMB000000144	48939649	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063383	PGSC0003DMB000000144	48939638	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063443	PGSC0003DMB000000144	48936605	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0063472	PGSC0003DMB000000144	48936336	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096176	PGSC0003DMB000000274	50099359	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096209	PGSC0003DMB000000274	50099766	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096224	PGSC0003DMB000000274	50149326	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096226	PGSC0003DMB000000274	50149452	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096243	PGSC0003DMB000000274	50149719	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096248	PGSC0003DMB000000274	50149787	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096290	PGSC0003DMB000000274	50592511	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096343	PGSC0003DMB000000274	50593073	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096344	PGSC0003DMB000000274	50593076	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096362	PGSC0003DMB000000274	50593176	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096412	PGSC0003DMB000000274	50594389	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096428	PGSC0003DMB000000274	50594858	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0096438	PGSC0003DMB000000274	50595039	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100103	PGSC0003DMB000000292	51136764	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100192	PGSC0003DMB000000292	51137938	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100194	PGSC0003DMB000000292	51137976	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100196	PGSC0003DMB000000292	51137984	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100264	PGSC0003DMB000000292	51252161	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100303	PGSC0003DMB000000292	51253527	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100405	PGSC0003DMB000000292	51488011	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100427	PGSC0003DMB000000292	51488249	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0100473	PGSC0003DMB000000292	51489509	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0081221	PGSC0003DMB000000201	52158070	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0081222	PGSC0003DMB000000201	52158074	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0081224	PGSC0003DMB000000201	52158082	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0081225	PGSC0003DMB000000201	52158088	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0081271	PGSC0003DMB000000201	52158707	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0081279	PGSC0003DMB000000201	52158803	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0119078	PGSC0003DMB000000445	52158899	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0119088	PGSC0003DMB000000445	53457373	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0119091	PGSC0003DMB000000445	53457295	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1
chr08	PotVar0119152	PGSC0003DMB000000445	53138672	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	0	0	1

Fig. S3.

Graphical genotyping of Kartel introgression segment identified in a variety panel of 83 varieties. This 5.8 Mb segment is underlying the Sgr7.1 QTL in the A x C population.

Chapter 5

Graphical genotyping as a method to map *Ry$_{sto}$* and *Gpa5* using a panel of tetraploid potato varieties

Herman J. van Eck[1,3], Peter G. Vos[1,3], Jan G.A.M.L. Uitdewilligen[1], Hellen Lensing[2], Nick de Vetten[2], Richard Visser[1,3]

[1] Plant Breeding, Wageningen University & Research , P.O. Box 386, 6700 AJ Wageningen, The Netherlands
[2] Averis Seeds B.V., Valtherblokken Zuid 40, 7876 TC Valthermond, The Netherlands
[3] Centre for BioSystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands

Submitted to Theoretical and Applied Genetics

Abstract

Graphical genotyping is a visually attractive and easily interpretable method to represent genetic marker data. In this paper the method is extended from diploids to a panel of tetraploid potato varieties. Application of filters to select a subset of the SNPs allows one to visualize haplotype sharing between individuals that also share a specific locus. The method is illustrated with two varieties SANTÉ and FESTIEN, both resistant to *Potato Virus Y*, while simultaneously selecting for the absence of the SNPs in susceptible clones. SNP data will then merge into an image that displays the coordinates of a distal genomic region on the north arm of chromosome *11* where a specific haplotype is introgressed from *S. stoloniferum* CPC 2093 carrying the Ry_{sto} gene. Graphical genotyping was also successful in representing the haplotypes on chromosome *12* carrying $Ry\text{-}f_{sto}$, as well as an image of chromosome *5* haplotypes from *S. vernei,* with the *Gpa5* locus involved in resistance against *Globodera pallida* cyst nematodes. The image also shows shortening of linkage drag by meiotic recombination of the introgression segment in more recent breeding material. The potential and limitations of the method are discussed and we show that IBD (identity-by-descent) is a requirement. Graphical genotyping is proposed as a non-statistical alternative compared to genome wide association studies (GWAS).

Introduction

Graphical genotyping

The concept of graphical genotypes as a visual method to represent genetic marker data was first described by Young and Tanksley (1989). In a karyotype-style drawing, they used colours to display the mosaic structure of the chromosomes in F_2 or backcross individuals according to the parental origin of the chromosomal regions or alleles. The transition of one colour into an alternative colour displays how chromosomes were transmitted from the parents, indicating the positions of crossovers that occurred during the meiosis of the F_1 plants. Graphical genotypes are visually attractive and in a glance more easily interpretable than the numerical information in a spreadsheet with offspring genotypes in columns and marker data in rows.

Graphical genotypes are used for a number of analyses. Firstly, graphical genotypes can display the positions and the proportions of donor and recurrent genome during subsequent backcross generations (Young and Tanksley, 1989; Yun et al. 2006). Secondly, graphical genotypes allow quick data inspection and the visual identification of singletons (**Fig. 1a**) that should not be interpreted as double recombinants (Van Os et al. 2005). Good data quality should result in a striping pattern whereas error adds isolated spots. Thirdly, high resolution mapping of major-effect loci (Finkers-Tomczak et al. 2009) is supported by a graphical analysis of the recombinants and their trait values, allowing one to zoom in on the remaining interval where candidate genes reside.

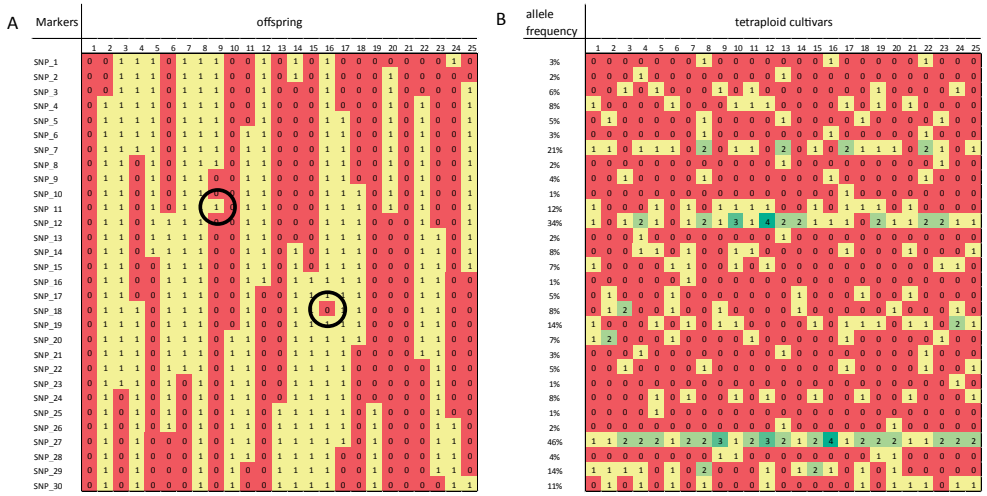

Fig. 1

Two hypothetical examples of graphical genotypes. **Panel A:** Graphical genotypes of a bi-parental diploid backcross offspring where colour codes distinguish the parental origin of the chromosomal segment (Red indicates the recurrent parent, Yellow indicates heterozygous introgression of donor parent). Singletons, putatively indicating error data, are indicated with the black circles. **Panel B:** Graphical genotypes of unrelated tetraploid varieties, where the dosage of the minor allele is indicated with numbers 0, 1, 2, 3 and 4 or the shades red, yellow, light green, middle green and dark green.

Graphical genotyping is supported by software packages such as GGT (van Berloo, 1999; 2008) or Flapjack (Milne et al. 2010).

Usually, graphical genotypes are generated using offspring from bi-parental crosses where two colours can encode the contribution of each of the two parents. A third colour may be needed to depict heterozygosity, but in BC_1 or RILs a third colour is not necessary. The F_1 offspring from non-inbred parents is already cumbersome, because four colours are required to show the mosaic structure in the maternally and paternally inherited recombinant chromosomes. This is easily circumvented by the separation of maternal and paternal markers and to graphically genotype the parental maps. An example is shown in **Fig. 1a**, which could represent a parental map of non-inbred parents, DHs, RILs or a BC_1.
Graphical genotyping of unrelated clones may still be possible for diploid selfing species. Here the graphical image can be used to display a local haplotype map (Motte et al. 2014) where DNA sequence variants representing either the allele of the reference genome or the non-reference allele are shown with different colour codes. Graphical genotypes or local haplotype maps are seemingly different terms (used in a mapping or haplotyping context, respectively) but both are conceptually related. Multi-colour graphical genotyping is used for depicting Multi-parent Advanced Generation Inter-Cross (MAGIC) lines (Bergelson and Roux, 2010). Graphical genotyping using unrelated tetraploids does not immediately display a meaningful pattern as shown in **Fig. 1b**, but when data are filtered valuable information can be retrieved.

PVY resistance
Potato virus Y (PVY) is a serious disease causing 50-85% yield reduction and economic losses during cultivation and the propagation of certified seed tuber material (Valkonen, 2007). Although many countries have implemented systems for the propagation of virus-free planting material, breeding for PVY resistance offers a more durable solution. Resistance genes against PVY have been introgressed in potato varieties from many sources as summarized in **Table 1**.

Solanum stoloniferum is an important source of PVY resistance, and the *Ry-f*$_{sto}$ and *Ry*$_{sto}$ genes on chromosome *12* have been used (Flis et al. 2005; Song et al. 2008). However, Brigneti et al. (1997) reported earlier than Song et al. (2005) on the localisation of a PVY resistance from *S. stoloniferum* on chromosome *11*, but their work has not received much follow up and their locus name *Ry*$_{sto}$ has remained confusing since then. Brigneti et al. (1997) used clone I-1039, which was developed in India as cv. KHUMAL RED 2 from Scottish late blight resistant progenitor clones and an undisclosed progenitor M 136-6. This material has *Solanum phureja* and *S. edinense* in its pedigree, but its *S. stoloniferum* origin is not understood from pedigree information by the authors. Clone I-1039 is

Table 1

Overview of the genetic loci involved in PVY resistance with their linkage group and ancestral germplasm

locus name	linkage group	ancestral germplasm	Reference
Nc_{spl}	4	*S. sparsipilum*	Moury et al. 2011
Ny_{tbr}	4	*S. tuberosum*	Celebi-Toprak et al. 2002
Ry_{chc}	9	*S. chacoense*	Hosaka et al. 2001; Sato et al. 2006
Ny-1	9	cv. Rywal	Szajko et al. 2008
Ny-Smira	9	cv. Sarpo Mira	Tomczyńska et al. 2014
Ry_{adg}	11	*S. tuberosum* Group *andigena*	Hämäläinen et al. 1997
Ny-2	11	cv. Romula	Szajko et al. 2014
Ry_{sto}	11	*S. stoloniferum*	Brigneti et al. 1997; this paper
$Ry\text{-}f_{sto}$	12	*S. stoloniferum*	Flis et al. 2005; Song et al. 2005

rarely used in breeding and resulted in two Danish varieties TIVOLI and LIVA, and two varieties released in Ecuador and Rwanda, FRIPAPA 99 and GIKUNGU, respectively. Hence, it may be problematic to extrapolate the marker data of Brigneti et al. (1997), for the identification of PVY resistance beyond their experimental mapping population. Another donor of PVY resistance, clone F87084, derived from *S. stoloniferum* CPC 2093 (De Jong et al. 2001; Nie et al. 2014) has not yet resulted in varieties released to the market and its map position is not described.

Another PVY resistance from *S. stoloniferum* origin was introgressed via three progenitor clones: MPI 13128 (sto x ERICA), clone 43 (sto x POLONIA), MPI 46.152/1 (sto x FRÜHMOLLE) (Song et al. 2008). Therefore it was obvious that research based in Germany and Poland, using predominantly German and Polish varieties could not reproduce the work by Brigneti et al. (1997). Their work casted doubt on either the localisation of Ry_{sto} on chromosome *11* or the identity of the source material that was used. It is now commonly accepted that Ry_{sto} and $Ry\text{-}f_{sto}$ are located on chromosome *12*. The Dutch variety SANTÉ is also PVY resistant and according to the pedigree database (Berloo et al. 2007) is also descending from *S. stoloniferum*. In this case the accession is known to be CPC 2093. Dutch potato breeders have informed the authors that markers developed by Song et al. (2005, 2008) or Flis et al. (2005) are not successful to identify the PVY resistance of this clone and other contemporary Dutch PVY resistant varieties. This failure to monitor the PVY resistance of cv. SANTÉ was also described by Heldák et al. (2007).

This study

In this paper we show that Dutch potato varieties descending via progenitor clone Y 66-13-636 from *S. stoloniferum* CPC 2093 carry a PVY resistance on chromosome *11*, which is not associated with male sterility, which causes a hypersensitive response (HR) to PVY.

This paper explores the ability to apply graphical genotyping on unrelated tetraploid varieties as a tool for identification of haplotypes and a tool for mapping PVY resistance. Furthermore we show that this method also enables the identification of introgression segments of the *Gpa5* locus (Rouppe van der Voort et al. 2000) and the *Ry-f*$_{sto}$ locus (Song et al. 2005; Flis et al. 2005). However, we also show that other traits without a clear most recent common ancestor (MRCA) are beyond the potential of this method, because identity-by-descent (IBD) is a crucial requirement of this method.

Materials and Methods

Plant material and PVY resistance data

The panel of 83 tetraploid varieties and progenitor clones is provided as **Table S1** described in Uitdewilligen et al. (2013). Among these tetraploids the authors were aware that variety FESTIEN is resistant to PVY. This is potentially due to a resistance gene derived from its great-grandparent Y 66-13-636 which is also included in the panel. SANTÉ was not included. KARTEL was included in the panel and this variety is the PVY susceptible parent of FESTIEN.

Validation of the PVY resistance gene was achieved by growing 180 offspring in pots in the greenhouse of the cross between FESTIEN (PVY resistant) ′ SERESTA (PVY susceptible).

Potato virus Y inoculation was performed after 3-4 weeks using mechanical inoculation using a field isolate of PVY^{NTN}. Two weeks later symptoms of primary infection were scored. All 180 genotypes were analysed by ELISA using monoclonal antiserum (obtained from Plant Research International, Wageningen, the Netherlands). Plants negative for PVY in ELISA and without visible symptoms were defined as resistant, whereas the remainder was defined as susceptible.

DNA sequence variants

The sequence of the reference genome of *Solanum tuberosum* Group Phureja DM1-3 516R44 (DM) was obtained from the Potato Genome Sequencing Consortium (PGSC, 2012). The order of superscaffolds (referred to as DMB###) is according to version 4.03 and can be retrieved from Sharma et al. (2013). Next generation sequencing of a panel of 83 tetraploid varieties allowed the identification of 129,156 DNA sequence variants (Uitdewilligen et al. 2013). Here, the relative read depth at variant positions was used for tetraploid genotype calling in the varieties. Further details about the methods that resulted in this dataset can be retrieved from Uitdewilligen et al. (2013). A DNA sequence variant from this study may refer to a SNP, a multinucleotide polymorphism (MNP), indels, and/or multi-allelic SNPs, hereafter for brevity collectively referred to as SNPs.

Genotype calling can be performed using the following terms: the reference (DM) and alternative (non-DM) allele, or the minor and major allele, abbreviated hereafter as REF and ALT or MIN and MAJ. Here, genotype calls representing the ALT dosage values indicate that a genotype has 0, 1, 2, 3 or 4 copies of the non-DM allele as compared to the DM reference (REF) genome. The variety panel of 83 tetraploids represents a population of 332 alleles (4*83) and for each DNA variant its allele frequency has been calculated as the sum of the copies per variety divided by 332. Hence, population allele counts ranging from 1 to 166 results in a minor allele frequency (MAF), and the allele with such an allele frequency equal or below 50% is considered as the minor allele

(MIN). Allele counts ranging from 167 to 331 result in an allele frequency >50% and such alleles are regarded as the major allele (MAJ).
Please note that colour codes are not defined relative to the DM reference genome. If the DM reference genome has a sequence variant, which is quite rare in the remainder of the gene pool then the DM reference genome represents the minor allele.

Filtering of SNPs dosage data to construct graphical genotypes.
The data were loaded in Microsoft Excel where the 83 varieties are shown in columns and the sequence variants along with their coordinates in 129,156 rows. The rows were sorted according to chromosome, superscaffold order and coordinates on pseudomolecules version 4.03 (Sharma et al. 2013). Filters were used in Excel to select chromosomal regions according to phenotypic criteria: e.g. specific minor alleles should be present in FESTIEN and Y 66-13-636. For columns representing susceptible varieties the filter was set at zero for the minor allele, which results in nulliplex genotypes. Graphical genotypes were displayed in colour using conditional formats in Excel where pseudo-colours indicated allele dosage per genotype ranging from red to green for nulliplex to quadruplex genotypes respectively.

GWAS
The above analysis was verified with a statistical analysis. For this purpose trait values were defined for 83 potato varieties (trait values are resistant = 1; susceptible = 0) and a naive genome wide association study (GWAS) was performed using GenStat 15th edition (VSN International Ltd., UK).

Results

When SNP data were loaded in Excel, conditional formatting according to colour scales were set as shown in **Fig. 1**. This resulted in a predominant background colour with rows containing fewer or greater numbers of cells having the alternative pseudo-colour according to the population allele frequency. This allows the interpretation that the vast majority of the SNPs alleles represent rare haplotypes. **Fig. 2** displays a clear L-shaped distribution of the allele frequencies of the SNPs. This implies that if SNPs were phased into haplotypes, the haplotype allele frequency will also display such an L-shaped distribution, which suggests that the potato gene pool is comprised of many low frequent haplotypes.

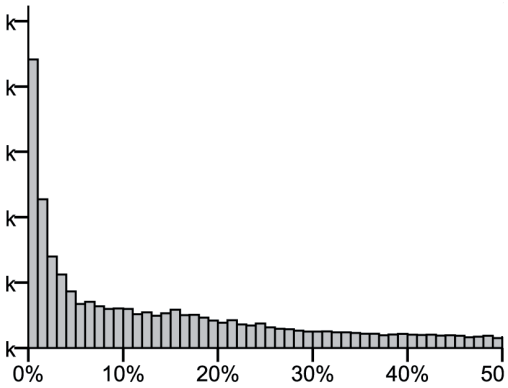

Fig. 2

Distribution of minor allele frequencies (MAF in y-axis) of all 129,156 genotyped sequence variants (after Uitdewilligen et al. 2013)

Filtering of SNPs data was performed with three criteria. Firstly, the PVY resistance has a low allele frequency, so the filter for the column with the allele count was set below 20 (allele frequency of 6.0%). Secondly, the filter of the two resistant clones FESTIEN and Y 66-13-636 were set to >1 because these varieties carry at least one allele conferring PVY resistance. Thirdly, for some susceptible clones such as KARTEL (the susceptible parent of FESTIEN), as well as KATAHDIN, BINTJE and ARRAN PILOT (old varieties) the filters were set to 0, indicating the absence of an allele involved in resistance.

This filtering resulted in only 531 SNPs matching these criteria, which represents only 0.41% of all SNPs. Interestingly, with conditional formats this SNP data produced a specific striping pattern as shown in **Fig. 3**. This pattern is indeed reminiscent of graphical genotypes as observed for diploid bi-parental F_2 or backcross populations. However, the variety panel is comprised of highly diverse material. If any haplotype would be shared between members of the variety panel, then haplotype specific SNP alleles (hs-SNPs) will merge into a yellow vertical bar in the graphical genotyping image. Haplotype blocks could be identified for regions that belong to potato chromosome *4*, *5*, *7*, *10* and *11*. For all regions, except the region on chromosome *11*, the yellow patterns are also observed for potato varieties that are susceptible. Based on the many false positives these regions were rejected. The candidate region on chromosome *11* indicated that one haploblock was shared by three varieties: EOS, FESTIEN and Y 66-13-636. This prompted us to examine the phenotype and pedigree of EOS. Indeed EOS appeared to be highly resistant against PVY, male fertile (Jacob Eising, personal communication, potato breeder at Den Hartigh b.v.), and has Y 66-13-636 as grandfather. Therefore, EOS is not a false positive but a third member of the variety panel with PVY resistance derived from *S. stoloniferum*.

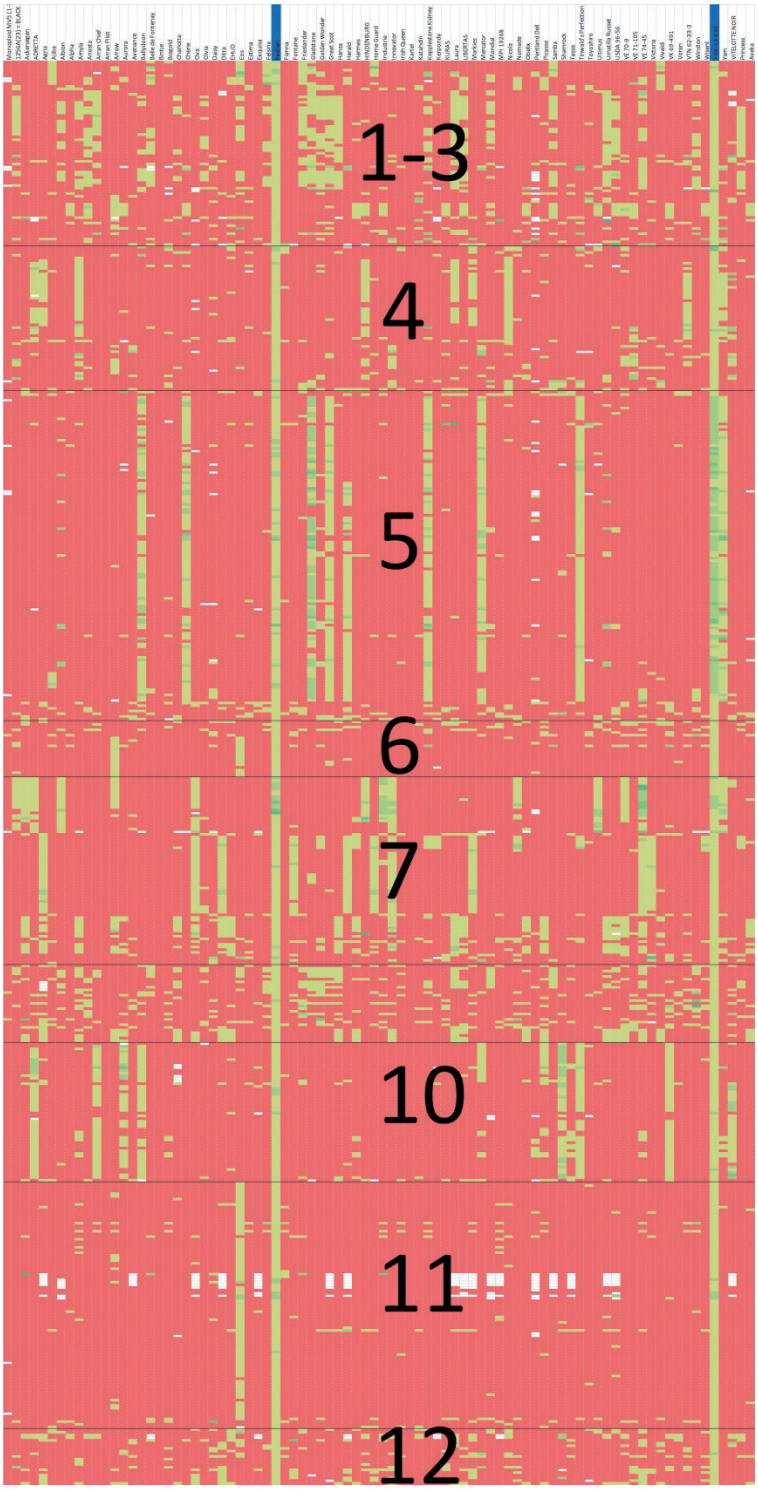

Fig. 3

Graphical genotyping of 83 tetraploid varieties (columns) with 531 SNPs (in rows) from 12 potato chromosomes. Absence or presence of minor alleles is indicated by conditionally formatted cells in red (0) or yellow to green (1, 2, 3, 4) respectively. Settings for allele dosages in FESTIEN and Y 66-13-636 were ≥1 (yellow to green columns, depending on allele dosage), whereas KARTEL, KATAHDIN, BINTJE and ARRAN PILOT were set <1 (red columns). The apparent pattern is indicative of haploblock sharing. The haploblock of chromosome *11* indicates the introgression segment that carries the PVY resistance gene descending from *S. stoloniferum* CPC 2093.

Size of the haploblocks

The map location and sizes of the haploblocks on potato chromosome *4*, *5*, *6, 7*, *10* and *11*, as shown in **Fig. 3**, can be deduced from the physical coordinates of the first and last SNP in the block. For chromosome *4* the haploblock is part of the pericentromeric heterochromatin and is roughly 10Mb long, ranging from DMB366 to DMB13 (PGSC 4.03 coordinates chr04:21464000…31110000). For chromosome *5* the haploblock is part of the north arm and is only 1.1 Mb long, ranging from DMB51 to DMB424 (PGSC 4.03 coordinates chr05:4250427…5369289). For chromosome *6* the 12 SNP haploblock locates on de distal DMB686 chr06:57791621…57992915. For chromosome *7* all SNPs belong to two scaffolds, both on the short arm. One block is part of the 2.2 Mb most-distal scaffold DMB47 and one block is 8 Mb further on the 0.2Mb scaffold DMB684. For chromosome *10* the 1 Mb haploblock on the north arm is part of two adjacent scaffolds DMB599 and DMB338 (PGSC 4.03 coordinates chr10:3903000…4917000). For chromosome *11* all SNPs of the haploblock belong to the most distal scaffold DMB148 on the north arm (PGSC 4.03 coordinates chr11:1…1439384). The first and the last SNP of the haploblock are 500 kb apart, having coordinates chr11:284162 and chr:814554, respectively. This is approximately 500 kb from the well-known *R*-gene cluster that includes genes homologous to *N* and *N*-like (*Nl-25*) TMV resistance genes and analogues (Hehl et al. 1999).

The haploblock comprising Ry_{sto} from S. stoloniferum CPC 2093

Further refinement of the chromosome *11* haploblock (**Fig. 3**) comprising Ry_{sto} from *S. stoloniferum* CPC 2093 was performed to identify accurate hs-SNPs using renewed filter settings, where the allele count was set to ≤4 (allowing at most one genotyping error), and EOS, FESTIEN and Y 66-13-636 have an allele count ≥ 1. This resulted in a total of 65 SNPs (**Fig. S1**) that displayed the minor allele in simplex condition in EOS, FESTIEN and Y 66-13-636 and was nulliplex in the other members of the variety panel. These SNPs should allow marker-assisted selection in any genetic background without the risk of false positive results. From these 65 SNPs a subset of seven SNPs was selected for validation. Their identity and coordinates are as follows:
PotVar0063974 (Chr11:284168) a [T/C] SNP, PotVar0064044 (Chr11:398248) a [T/C] SNP, PotVar0064080 (Chr11:398597..398598) a dinucleotide polymorphism [TT/GA], PotVar0064400 (Chr11:786626) a [G/A] SNP, PotVar0064470 (Chr11:787325) a [C/A] SNP, PotVar0064502 (Chr11:787571) a [C/T] SNP, and PotVar0064578 (Chr11:809990) a [T/C] SNP, all from DMB148. The mapping of Ry_{sto} to DMB148 suggests that the R-gene is a member of a well-known R-gene cluster, which was named cluster XIa-TNL in bin4-8 (Bakker et al. 2011), or the cluster 49 (Jupe et al. 2012). The seven SNPs were selected to be free (as much as possible) from flanking SNPs to avoid assay failure, and were included in a custom-made 20K Infinium SNP array (Vos et al. 2015) used to analyse a much wider panel of 537 potato varieties described by

D'hoop et al. (2014). In this panel the seven SNPs were validated, because all SNPs positively identified other descendants from Y 66-13-636: CUPIDO, CYCLOON, LADY CHRISTL, LADY FELICIA, MELODY, MUSICA, ORCHESTRA, SANTÉ, SAVIOLA, W 72-22-496 as well as MIRAKEL descending from Y 62-2-221 (the resistant parent of Y 66-13-636). These varieties are simplex for the minor allele and their PVY resistance is in agreement with phenotypic data from breeders' websites. In susceptible clones however, the SNPs identified only the major reference allele.
The Infinium array also enabled the identification of four varieties: ARIZONA, BELANA, OSIRA and SAGITTA which are positive for this introgression segment, but bear an unknown genetic relationship to CPC 2093. Hence we predict that these varieties are resistant, which was verified at the breeders' website. Furthermore, ALTUS, AXION, CYRANO, DONALD, SERESTA and XANTIA were negative for the SNPs, although these varieties descend from CPC 2093 (van Berloo et al. 2007). It is concluded that these clones no longer carry the introgression segment and are most likely PVY susceptible.

Validation of the map position with a bi-parental population

The Ry_{sto} locus identified using a panel of distantly related varieties was validated with a bi-parental mapping population descending from FESTIEN × SERESTA. In this mapping population, 81 individuals were found to be susceptible, whereas 98 were found resistant (**Fig. 4**). This is in agreement with a 1:1 ratio, in accordance with the Mendelian expectations of a simplex allele. The SNP marker PotVar0063973 (Chr11:284162), a [T/C] SNP, and PotVar0063974 (Chr11: 284168), a [T/C] SNP, were converted into one SNP assay and was analysed for co-segregation with the PVY resistance. The SNP assay nicely predicted the PVY resistance phenotype in the offspring of 179 individuals, except for four resistant individuals. This may indicate four recombination events between the R-gene and the marker locus PotVar0063973/0063974, but the resistant phenotype may also indicate an unsuccessful inoculation. The latter is more likely, because in the

Fig. 4

Observed phenotypes upon inoculation with PVY^{NTN} in FESTIEN × SERESTA offspring. A: A descendant conferring resistance to PVY, (B) and a susceptible descendant, two weeks after mechanical inoculation. (C) Leaves cut from a whole plant, to see the phenotypes from a resistant genotype (left) and susceptible genotype (right).

wider panel of 537 varieties tested with the Infinium array no recombination events were observed among the SNP loci. These results provide an independent confirmation of the localisation of PVY resistance at a distal position of potato chromosome *11*.

Results of GWAS

A genome wide association study was performed as a statistical alternative to the graphical genotyping method to analyse the significance of the mapping and the power of GWAS to identify the PVY resistance based on only four resistant varieties (EOS, FESTIEN, KURAS and Y 66-13-636) in the panel of 83 tetraploids. We included cv. KURAS as the fourth PVY resistant clone to mimic the situation that GWAS is based on phenotypic data and does not have a priori assumptions on trait heterogeneity (= Variation caused by different genes can give rise to the same phenotype). **Fig. 5** shows a Manhattan plot with the outcome of the GWAS for Y-virus resistance. A major peak can be observed for SNPs that reside in superscaffold DMB148 on chromosome *11*. These SNP explain up to 75% of the phenotypic variance. Additional (singleton) significant SNPs were detected on chromosomes *2*, *3*, *6*, *7* and *12*. The SNPs underlying the peak on the right end of chromosome *12* reside in superscaffold DMB114 and represent the haplotype that carries *Ry-f*$_{sto}$ (Song et al. 2008). The peak on the right end of chromosome *6* represents a false positive QTL, based on coincidence (see the graphical genotyping

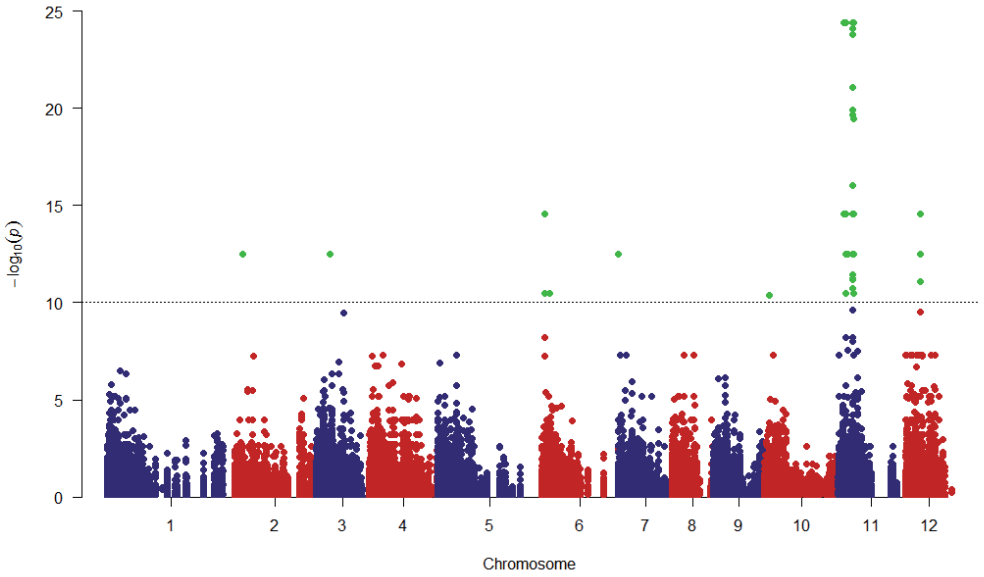

Fig. 5

Manhattan plot of a GWAS, showing the statistical validation of the graphical mapping of *Ry*$_{sto}$ on superscaffold DMB148 on potato chromosome *11*. A secondary peak on chromosome *12* is indicative of *Ry-f*$_{sto}$. On the x-axis the physical positions of the SNP is shown in alternating colours for twelve consecutive potato chromosomes.

image Fig.3). In a GWAS where KURAS is excluded from the analysis, the secondary peak on chromosome *12* is absent (data not shown).

Graphical genotyping of Ry-f$_{sto}$

Upon identification of *Ry$_{sto}$*, we aimed to map other traits with the method of graphical genotyping by imposing absence / presence criteria on the minor SNP allele with filters in Excel. We tested if graphical genotyping could identify the *Ry-f$_{sto}$* locus. In our variety panel only KURAS could be identified as being PVY resistant due to *Ry-f$_{sto}$*. A value for allele count was set <2, and a value for a minor allele ≥1 for KURAS. This results in the identification of each of the 479 KURAS-specific SNPs as potentially associated with the *Ry-f$_{sto}$* haplotype. However these SNPs were located on segments spanning the entire genome. Scaffold DMB127 on chromosome *3* contributed 75 SNPs (16%); twelve scaffolds of chromosome *7* contributed 321 SNPs (67%), and 50 SNPs (10%) localized to chromosome *12*. Within chromosome *12* the SNPs localized at three positions: 11 SNPs telomeric north arm, 13 SNPs close to the centromere and telomeric south we observed 2 SNPs in DMB38 and adjacently 24 SNPs in DMB114 (length 1.6 Mb , coordinates chr12:58927871…60583523). This is indeed the *R*-gene cluster tagged by STM0003 (coordinates: chr12:60055231…60055365; Song et al. 2005) and the PCR markers YES3-3A (genbank accession BV725480; BLAST hit on coordinates: chr12:59061649…59061903)

Nevertheless this result is incorrect. Three of the SNPs from this introgression segment were present on our 20K SNP-array, but were negative for AMADO, a PVY resistant descendant of KURAS that was present in the wider panel of 537 potato varieties. However, other SNPs on the Infinium array could accurately predict PVY resistance based on *Ry-f$_{sto}$*. Therefore graphical genotyping was performed with more relaxed settings. When the filter for allele count was set ≤ 5 an unexpected haplotype sharing between cv. KURAS and cv. HINDENBURG could be observed in the graphical genotyping image. From other data we already knew that our DNA was isolated from a clone incorrectly labelled as HINDENBURG.

A second attempt included this incorrectly labelled clone (presumably carrying *Ry-f$_{sto}$*) along with the additional knowledge on the location of *Ry-f$_{sto}$* on DMB114, chromosome *12*. This resulted in the identification of 62 SNPs (**Fig. S2**). Four of these 62 SNPs (PotVar0052353 at DMB114:626106, PotVar0052707 at DMB114:902759, PotVar0053235 at DMB114:1376622 and PotVar0053451 at DMB114:1599751) were present on the 20K Infinium array and all four SNPs positively identified KURAS and AMADO and were negative for the remainder of the 537 variety panel.

The Solanum vernei introgression segment with Gpa5

Another example of the application of graphical genotyping relates to the $Gpa5_{vrn}$ locus on chromosome *5* conferring *Globodera pallida* nematode resistance, initially described using the diploid clone 3704-76 (Rouppe van der Voort et al. 2000). The resistant parental clones 3778-16 and 3704-76 are dihaploids extracted from the tetraploid progenitor clones AM 78-3778 and AM 78-3704, respectively. Both clones may carry multiple *S. vernei* derived alleles from both the maternal and paternal side. The pedigrees of both AM 78-3778 and AM 78-3704 display the same three sources of *S. vernei* (van Berloo et al. 2007). The first is the breeding clone LGU 8 (developed by R.L. Plaisted, Cornell) using unnamed *S. vernei* material from Hans Ross (MPIZ, Germany). The second is clone V 24/20 which points back to Scottish work using CPC 2488-3 x CPC 2487-3. And thirdly VRN 1-3 represents material derived from *S. vernei* hybrid GLKS 58.1642-4 (Dellaert and Vinke 1987).

Irrespective of complex ancestries, it is commonly known that a variety such as INNOVATOR has an exceptionally high level of resistance, which is derived from progenitor AM 78-3778.

Hence we filtered on INNOVATOR as the resistant clone, selected SNPs from chromosome 5 with an allele count <20, and used ARRAN PILOT, BINTJE, CHARLOTTE, CIVA, DAISY, GOLDEN WONDER, HOME GUARD, KATAHDIN, ULTIMUS and YAM as negatives. This resulted in 294 remaining SNPs according to the pattern shown in **Fig. 6**.

A second remarkable observation from the graphical genotyping image shown in **Fig. 6** is the potentially erroneous orientation of DMB103 and the historical recombination events, which caused a reduction of the linkage drag associated with *Gpa5* introgression. The varieties AVENANCE (2005), INNOVATOR (1999) and WISENT (2005) have the largest introgression segment, followed by NOMADE (1995). In MERCATOR (1999) and KARTEL (1994) the introgression segment is again shortened and linkage drag is minimal in FESTIEN (2000). The shortening of the introgression segment (which is the same as the removal of linkage drag) does not follow a trend from older to more recently released varieties. One generation of clonal selection hardly fits within a decade of potato breeding. The minimal segment observed in cv. FESTIEN is consistent with the location of the HC marker (Achenbach et al. 2009). The HC marker is known to produce the least amount of false negatives of all markers developed to select *Gpa5*. Indeed, a BLAST search using the primer sequences of the HC-marker finds the best hit within DMB424 at superscaffold coordinates 174417…174688. Further analysis of haplotype sharing and inspection of pedigree information has allowed us to reconstruct the history of the inheritance and recombination of individual alleles. We conclude that progenitor clone AM 72-4454 is the most recent common ancestor of the material analysed here.

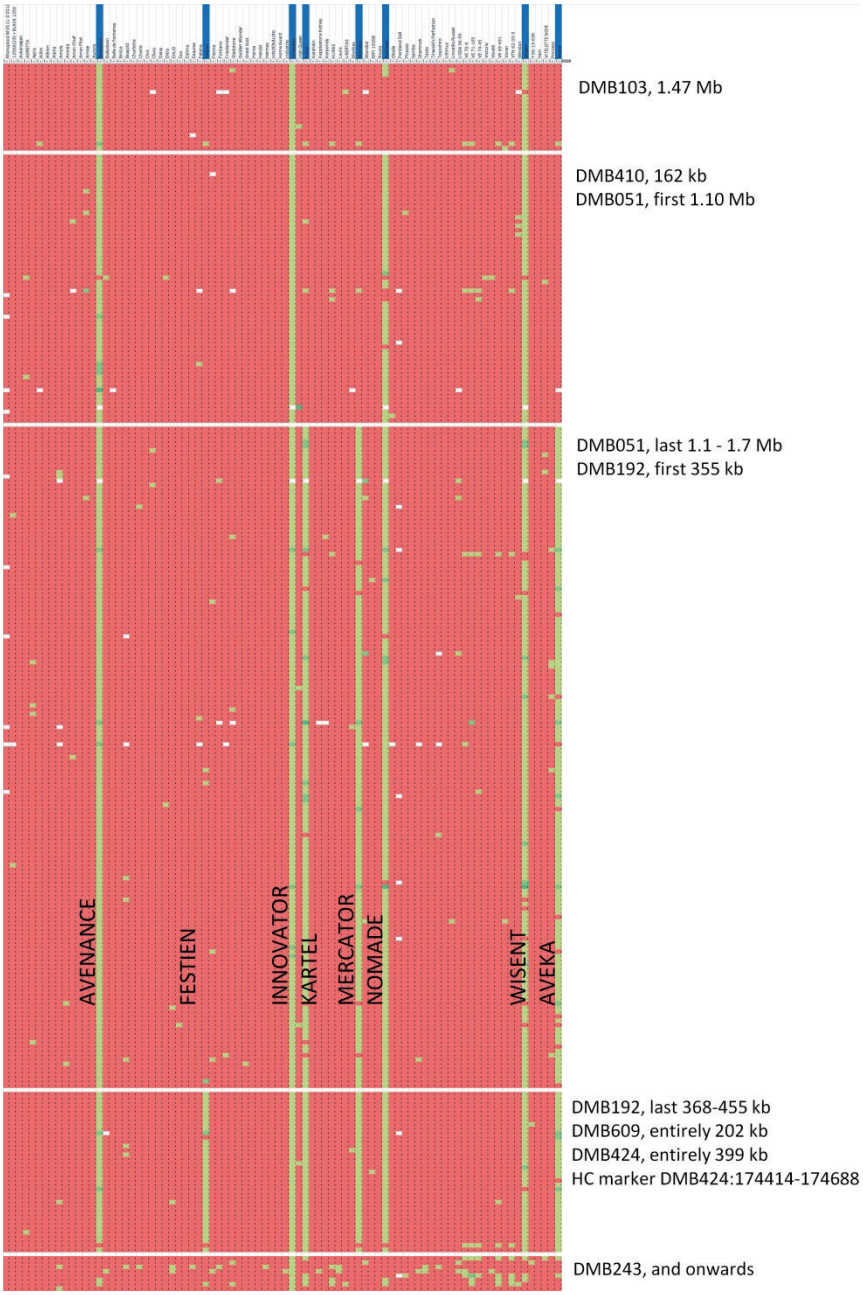

Fig. 6

Graphical genotyping image of SNPs from potato chromosome *5* where *Gpa5* haploblock sharing and size of linkage drag is visualized. The reference panel of 83 varieties (columns) has eight varieties with resistance against cyst nematodes *Globodera pallida* (race pa2/3). The introgressed haplotypes of AVENANCE, INNOVATOR and WISENT comprise the entire north arm of chromosome *5*. Variety FESTIEN has lost most linkage drag with an introgression segment of ~ 0.7 Mb.

Negative results

Unfortunately, we were unable to find SNPs indicating the *Gro1-4* gene on chromosome *7* involved in resistance against *Globodera rostochiensis* nematodes (also known as the *Fb* gene, Paal et al. 2004). This effort failed because it is unknown how the experimental material, from which the gene was cloned, is related to varieties that entered the market. Possibly the *Gro1-4* gene is hardly deployed and therefore escapes detection. Furthermore, *Ro5* resistant varieties descending from *S. spegazzinii* have an overly-complex pedigree resulting in ambiguity if they share any most recent common ancestor.

For the single locus traits flesh colour (Wolters et al. 2012), tuber shape (Van Eck et al 1994, 2014) and maturity (Kloosterman et al. 2013) this method failed. This is likely to be due to a violation of the assumption of identity-by-descent. If a phenotype is not the result of a single (recent) mutation, also indicated as genetic heterogeneity, then other (statistical) methods will also fail. In addition, the size of the haploblock may be too short to allow the identification of a meaningful tract of SNPs in a graphical genotyping image. Finally, it should be noted that although haplotypes are defined by SNPs, it is not necessarily so that any of these SNPs are haplotype specific. We therefore conclude that the power of graphical genotyping and the power of statistical methods equally suffer from genetic heterogeneity.

Discussion

Intuitively one may easily accept that application of a few selection criteria imposed on a data set with 129,156 SNPs allows the graphical visualisation of haploblocks. To explain the underlying principles is however not as easy. SNPs are binary characters that allow one to separate haplotypes into two groups: a group with haplotypes that share the SNP allele and a group of haplotypes that share the alternative variant. The number of haplotypes in the potato gene pool is however much larger and can be deduced from the large number of SNPs with a low minor allele frequency (MAF) as shown in **Fig. 2** with an average MAF of 14% only. Moreover, 17.4% of all SNPs have a MAF < 1%, and 39.1% have a MAF < 5% (Uitdewilligen et al. 2013). Supposing that all SNPs with a MAF < 5% are haplotype specific (hs-SNPs), this allows one to imagine that the gene pool is comprised of at least a dozen of different haplotypes per locus. Apart from the hs-SNPs, the remaining non-haplotype specific SNPs with a higher MAF may consign haplotypes into two groups, without any reason to assume that the alleles within such a group have anything in common.

Graphical genotyping thus relies on the ability to distinguish specific minor alleles because of the presence of hs-SNPs against the many other haplotypes that share the

absence of this minor hs-SNP allele. In other words, haplotype specificity refers to the minor allele - the major allele is not haplotype specific at all.
Another notion to understand this application of graphical genotyping in unrelated tetraploids is based on the breeding history of potato varieties. During a century of potato breeding and approximately one meiosis per decade (Love 1999), large haploblocks can be expected. The gene pool has a limited number of founders and donors of resistance and therefore identity-by-descent is obvious for haplotypes shared by related varieties. These aspects contribute to long haploblocks. Shorter haploblocks associated with traits can also appear by chance. The probability for such random effects is also influenced by the nucleotide diversity between haplotypes. .
Finally, within a given level of LD, nucleotide diversity and founders, the number of SNPs needs to exceed a certain minimal number. Only with a large number of SNPs clear stripes will appear within the image, where the length of the stripes enables the visual identification of haplotype sharing.
We have mapped a locus involved in PVY resistance derived from CPC 2093. Our data suggest a rehabilitation of the paper by Brigneti et al (1997). We propose to use the original locus name Ry_{sto} to indicate the PVY resistance gene on chromosome *11* that is not associated with male sterility. The locus name $Ry\text{-}f_{sto}$ can be used to indicate the PVY resistance on chromosome 12 that is found in the fertile parent PW-363 and many German varieties that are usually male sterile. The material used by Brigneti et al (1997) and the widely used resistance donor Y 66-13-636 could well be identical-by-descent. This is supported by presence the SNP markers in the Danish variety TIVOLI, a first generation descendant of the resistant parent I 1039 used in Brigenti et al (1997). Our SNP markers allow marker-assisted breeding and validation with an Infinium SNP array on a wider panel of 537 varieties was flawless.
We conclude that graphical genotyping is not only suitable to map loci in bi-parental mapping populations, but also in panels of distantly-related varieties. The graphical genotyping patterns observed here suggest the presence of specific haplotypes which are uniquely tagged by tracts of haplotype specific SNPs.
Graphical genotyping was initially proposed as a tool in mapping studies, but it also makes singleton observations visible to allow correction of erroneous genotyping results. Whereas the data structure from a panel of varieties (as shown in **Fig. 1B**) suggests that correction of genotyping errors, or the identification of mislabelled of plant material is not straightforward, this paper shows that skilful use of graphical genotyping images in variety panels can assist in making various interpretations of the data, including observation of trait heterogeneity and genetic heterogeneity.

Author Contribution

Conceived and designed the experiments: HvE. Performed the experiments: JU, HL, NdV. Analysed the data: HvE, JU, PV. Wrote the paper: HvE, PV.

Acknowledgements

This research was partially supported by a grant from the Dutch technology foundation STW, project WPB-7926. Peter Vos is supported by a grant from potato breeding companies Agrico Research B.V., Averis Seeds B.V., HZPC Holland B.V., KWS POTATO B.V. and Meijer B.V. We thank Peter Bourke for critical reading of the manuscript.

Table S1

List of varieties

Variety	Year of release	Country of origin	Variety	Year of release	Country of origin
1256A(23)	1945	GB	Innovator	1999	HOL
Ackersegen	1929	GER	Irish Queen	1900	GB
Adretta	1975	GER	Kartel	1994	HOL
Agria	1985	GER	Katahdin	1932	USA
Ajiba	1992	HOL	Kepplestone Kidney	1900	GB
Albion	1895	HOL	Kerpondy	1949	FRA
Alpha	1925	HOL	Kuras	1996	HOL
Amyla	1999	FRA	Laura	1998	GER
Anosta	1975	HOL	Libertas	1946	HOL
Arran Chief	1911	GB	Markies	1997	HOL
Arran Pilot	1930	GB	Mercator	1999	HOL
Arrow	2004	HOL	Mondial	1987	HOL
Aurora	1972	HOL	MPI 19268	1945	GER
Avenance	2005	HOL	Nicola	1973	GER
Ballydoon	1931	GB	Nomade	1995	HOL
Belle de Fontenay	1885	FRA	Obelix	1988	HOL
Bintje	1910	HOL	Pentland Dell	1961	GB
Biogold	2004	HOL	Picasso	1994	HOL
Charlotte	1981	FRA	Samba	1989	FRA
Cherie	1997	FRA	Shamrock	1900	IRL
Civa	1960	HOL	Tasso	1963	GER
Clivia	1962	GER	Tinwald's Perfection	1914	GB
Daisy	1998	FRA	Toyoshiro	1976	JPN
Ditta	1989	AUT	Ultimus	1935	HOL
Ehud	1965	HOL	Umatilla Russet	1998	USA
Eos	2000	HOL	USDA 96-56	1945	USA
Estima	1973	HOL	VE 70-9	1970	HOL
Exquisa	1992	GER	VE 71-105	1971	HOL
Felsina	1992	HOL	VE 74-45	1974	HOL
Festien	2000	HOL	Victoria	1997	HOL
Fianna	1987	HOL	Vivaldi	1998	HOL
Fontane	1999	HOL	VK 69-491	1969	HOL
Frieslander	1990	HOL	Voran	1931	GER
Gladstone	1932	GB	VTN 62-33-3	1962	HOL
Golden Wonder	1906	GB	Winston	1992	GB
Great Scot	1909	GB	Wisent	2005	HOL
Hansa	1957	GER	Y 66-13-636	1966	HOL
Herald	1928	GB	Yam	1787	GB
Hermes	1973	AUT	Vitelotte Noire	1815	FRA
Hindenburg	1916	GER	Princess	1998	BRD
Home Guard	1943	GB	Aveka	2001	HOL
Industrie	1900	GER			

Fig. S1. Graphical genotyping of the Ry_{sto} locus on chromosome 11.

Fig. S2 Graphical genotyping of the Ry-f_{sto} locus on chromosome 12.

Chapter 6

General discussion

Introduction

The research described in this thesis is a continuation of two earlier PhD thesis research projects. One carried out within the framework of the Centre of Biosystems and Genomics (CBSG), entitled "Association Mapping and Family Genotyping" aimed to explore the possibilities of GWAS in tetraploid potato with AFLP and SSR markers. A description of the statistical methods and the phenotypic data were presented along with the marker trait associations in a PhD thesis (D'hoop 2009) and several papers (D'hoop et al. 2008, 2010, 2011, 2014). Unfortunately, with today's knowledge it was naïve to expect that the marker density achieved with AFLP and SSR would be sufficient to allow detection of the most relevant QTLs. Furthermore the AFLP and SSR marker system is inconvenient to allow large-scale application of marker-assisted selection. On the other hand, AFLP was in those days one of the more efficient marker technologies. A second PhD project "Potatoes with novel properties for consumption and processing industry" (STW grant WPB-7926), resulted in an extremely valuable data-source of 129,156 DNA sequence variants from 83 potato genotypes (Uitdewilligen et al. 2013). The detection of highly significant marker-trait associations using a population of only 83 potato genotypes (Kloosterman et al. 2013; Uitdewilligen et al. 2013) suggested that GWAS would clearly benefit from additional SNP markers, and would significantly improve our ability to detect QTLs. Furthermore, new phenotypic data of breeding material was collected for subsequent validation studies. This would allow us to test the reproducibility of marker-trait association identified with GWAS. In this discussion I reflect on how the results described in experimental **chapters 2, 3, 4** and **5** have contributed to a better understanding of SNP-arrays, GWAS and LD-decay in tetraploid potato and I will discuss what implications these results have on the application of marker assisted selection/breeding (MAS/MAB).

As described in the general introduction genetic gain for yield in potato is limited, and that the application of markers assisted breeding/selection is still in its infancy. The use of molecular markers in potato is already a promising tool to assist in potato breeding for nearly three decades, however resources that can really boost marker-assisted selection in potato are only available for a relative short period. The most valuable resource that serves the scientific and breeding community is the reference genome of potato (PGSC 2011). This allows breeders and researchers to communicate on QTLs, candidate genes and genetic variants using the same reference coordinates for genetic loci and their SNP markers. In addition to that, vast numbers of natural sequence variants have been identified in two large re-sequencing efforts (Hamilton et al. 2011; Uitdewilligen et al. 2013). The combination of both the reference genome and the two collections of sequence variants allow researchers and breeders to develop mapping tools to identify QTLs, and understand from the underlying candidate genes their biological effect on their traits of interest. Above all, it allows the potato community to connect the

results of different studies. The SolCAP array (Felcher et al. 2012) was the first high-throughput genotyping array and it is widely used by the potato community. This array was based on sequence variants typically for North American processing varieties and therefore assumed to have limited value for European germplasm. The observation by Uitdewilligen et al. (2013) that 50% of all SNPs have an allele frequency below 5% suggested that many haplotypes could escape detection if such haplotype specific SNPs were omitted just on the basis of their low population allele frequency. Therefore, no allele frequency threshold was used when developing another SNP array, named the SolSTW array. This array contains 14,530 high quality SNPs and its development is described in **Chapter 2.** This SolSTW array is used to characterize a large set of potato varieties covering the worldwide gene pool of cultivated potato. And subsequently it was used to analyse linkage disequilibrium and population structure (both **Chapter 3**), in order to perform the main goal of this thesis: exploring the possibilities and pitfalls of genome wide association studies in tetraploid potato (**Chapter 4)**.

GWAS in potato

Genome wide association studies are a useful tool to unravel complex traits in crop species. In many crop species association studies have been performed and identified significant marker-trait associations. However, true genome wide association studies have not been conducted abundantly in potato. Besides the strengths of GWAS, the method suffers from several limitations as reviewed by Korte and Farlow (2013). Several major issues that play a role in the success or failure of GWAS such as population structure, ascertainment bias, reproducibility and marker density will be discussed below.

Population structure

Population structure can have a major effect on the results of an association study. In **Chapter 4** of this thesis we show that when a trait is confounded with population structure it can result in many spurious associations when a statistical correction for population structure is not applied. The correction for population structure with a kinship matrix may result in a substantial amount of missing heritability. In **Chapter 3** we analysed population structure, resulting in the identification of two "subpopulations" that represent specific breeding efforts for different market segments: starch varieties and French fry varieties, where the latter grouping is mainly caused by the frequent use of the variety Agria as parent. These two subpopulations were also identified with AFLP markers (D'hoop et al. 2010). Studies on North American material also show a separation of subgroups based on market segments (Hirsch et al. 2013; Rosyara et al. 2016). In contrast, in several German studies with undisclosed material from German

breeders hardly any structure was observed (Li et al. 2008; Li et al. 2010; Li et al. 2013; Schreiber et al. 2014; Schönhals et al. 2016).
Because the German contributions are contradictory with the US and NL papers, the issue of population structure should be reviewed at greater depth. The explained variance of principal coordinates/components allows to compare sub-population structure quantitatively between studies. In **Chapter 2** we report on a total of 6.7% explained by the two first principal coordinates, comparable to the value of 8% reported for US material (Rosyara et al. 2016). A substantial 12.7% was reported by Urbany et al. (2011), whereas only 2.1% was reported by Schönhals et al. (2016). In other studies no significant subpopulation structure was detected (Li et al. 2008; Li et al. 2010; Li et al. 2013; Schreiber et al. 2014).
Based on the relative small proportion of explained variance by the first two principal coordinates/components I conclude that potato has in general not a very clear separation between the different subgroups and the observation of sub-groups may depend on the choice of material. Nevertheless, in this thesis I have shown a major effect of population structure on the results of a genome wide association study (**Chapter 4**), because the trait values were highly confounded with the subpopulation structure. On the other hand I have shown that in spite of sub-populations correction did not affect the results of GWAS. In this case the trait values were not confounded with population structure (**Chapter 4**).
In this thesis a kinship matrix (K) was used in a mixed model to correct for population structure to avoid the detection of spurious associations (**Chapter 4**). Structure correction has an inherent negative consequence, because it may contribute to the phenomenon of missing heritability. For example the causal genetic variation for the high phenotypic values of total glycoalkaloid content that are confounded with population structure can't be identified using a kinship correction i.e. despite the kinship correction some true associations might not be identified.
To avoid the costs of kinship correction I propose an optimization of the composition of variety panels. In **Chapter 4** the variety panel was selected to perform a genome wide association mapping study for multiple traits (D'hoop et al. 2014), and therefore it was suboptimal for every single traits, such as total glycoalkaloid content. On the other hand the panel was suitable for analysing the ratio between **α**-solanine and **α**-chaconine; a trait that was never under selection by breeders. Those traits most eagerly selected by breeders are expected to be confounded with subpopulation structure. Ultimately, the recurrent use of successful varieties for specific markets as progenitors in breeding programs may reinforce sub-groups. To optimize an association panel several recommendations can be offered. Firstly population structure has to be taken in account beforehand. In our study we estimated population structure afterwards and ended up with small subpopulations without any statistical power on itself. To increase power within each subpopulation a denser sampling of genotypes in these subpopulations should be applied. Secondly

varieties or breeding clones within the subpopulations with non-preferable phenotypes should be included. For example we defined the "starch" subgroup in **Chapter 4** and one can imagine that all varieties in this subpopulation are high in starch. In this starch subpopulation the low starch genotypes are extremely valuable to circumvent the confounding of trait values with population structure and therewith gain power in detection of true QTLs. And finally an association panel should be composed with focus on a few traits maximally and with material that potentially can be used in a breeding program i.e. all old varieties and progenitors are most likely not used in crosses anymore and should therefore not be included in an optimal association panel. The older varieties were initially included to capture as much of the genetic variation as possible, this was proven unnecessary, because we show that the vast majority of the pre-1945 genetic variation remains present in new varieties (**Chapter 2**).

Ascertainment bias

The second factor that plays an important role in the success of GWAS is the amount of ascertainment bias. Ascertainment bias is defined as the inability to discover QTLs due to differences in genetic variation present in the population where the SNPs were discovered and the population where the SNP assays are applied. Consequences of disregarding ascertainment bias are well illustrated in rice, where different SNP arrays were tested on the rice gene pool (Thomson et al. 2012). Arrays were typically developed using SNPs from the only one gene pool (*indica*, *japonica*, *aromatic*, *aus*) and subsequently applied to other gene pools. Little variation could be discovered when using SNPs from a discovery panel that did not match the experimental material.
In this paragraph I wish to address to which level our research was negatively affected by ascertainment bias and to compare our situation with other published research efforts. In **Chapter 2** we have described the issue of ascertainment bias. We state that the wider discovery panel of Uitdewilligen et al. (2013) suffers less from ascertainment bias as compared to the SolCAP array, which is based on a small set of North American processing varieties (Hamilton et al. 2011). We also stated that the amount of ascertainment bias is difficult to estimate. However, in this concluding chapter I present new results, which allows me to address ascertainment bias. **Fig. 1a** shows a Manhattan plot using the VCU data (described in **Chapter 4)** for *Globodera pallida* pathotype 3. A clear peak on chromosome *5* is visible on exactly the same location where the *Gpa5* QTL and HC-marker were mapped in previous studies (Rouppe van der Voort et al. 2000; Sattarzadeh et al. 2006). The highly significant SNPs are the same SNPs identified with graphical genotyping in **Chapter 5** and these SNPs all are identified as originating from the same *Solanum vernei* progenitor used to introduce *Globodera pallida* resistance in commercial potato varieties (**Chapter 2)**. The most important observation to illustrate

ascertainment bias is the absence of highly significant SolCAP SNPs. The absence of significant SolCAP SNPs can be explained by the absence of *Globodera pallida* resistant varieties in the SNP detection study of Hamilton et al. (2011) i.e. ascertainment bias.

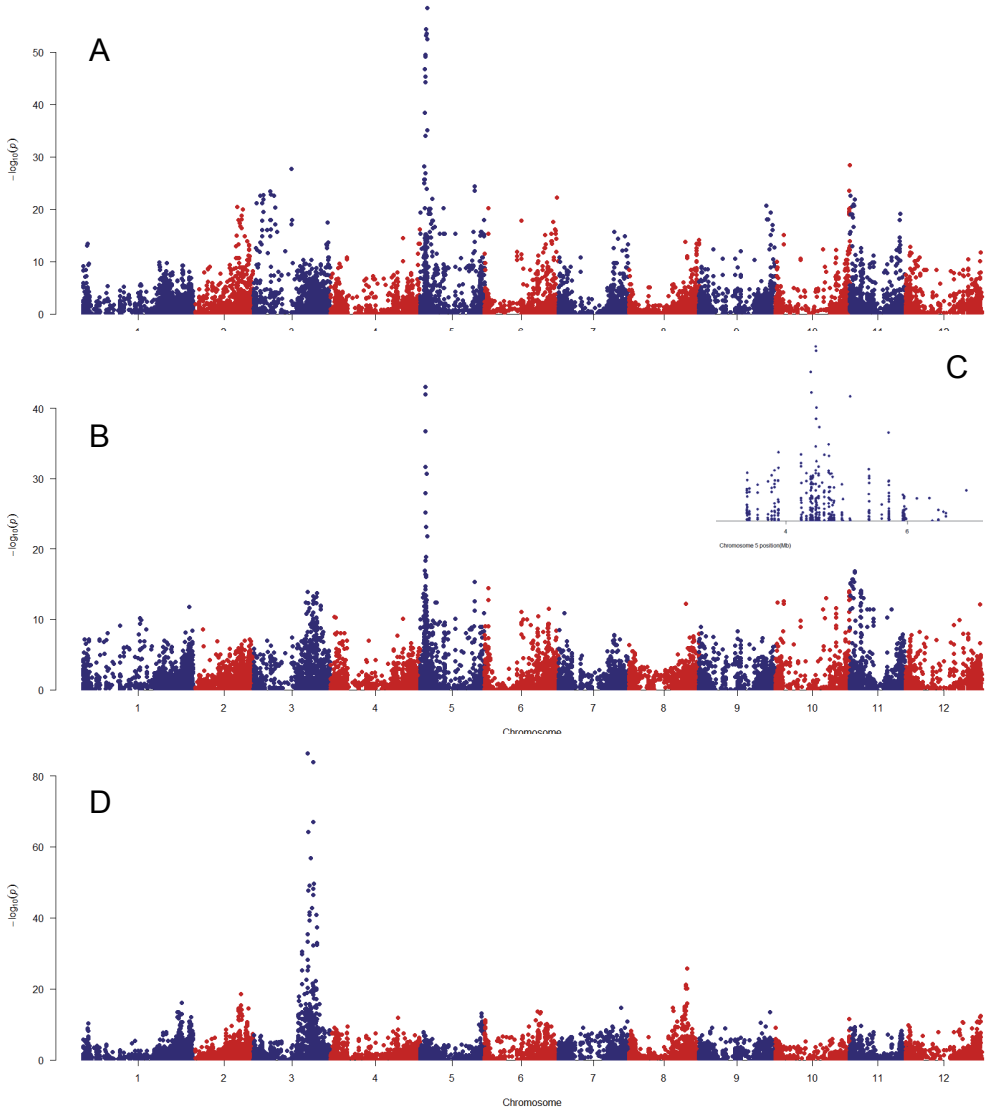

Fig. 1

Manhattan plots of naive GWAS analyses using VCU data for (A) resistance to *Globodera pallida* pathotype 3, (B) Foliage maturity, (C) Foliage maturity zoomed in around StCDF on chromosome *5* and (D) tuber flesh colour

The phenomenon ascertainment bias is also described by Obidiegwu et al. (2015). They state that none of the 6 varieties in the detection study (Hamilton et al. 2011) has resistance to wart disease (*Synchytrium endobioticum*) pathotypes 2, 6 and 18. Therefore, the identification of any major resistance gene is not expected. Nevertheless, they present many SolCAP SNPs with a "significant" association with resistance to all of the pathotypes of wart disease. In view of ascertainment bias these can't be the SNPs linked in coupling phase with the major resistance alleles. Putatively, these markers may have been successful to identify relevant loci in a bi-parental population because of linkage in repulsion phase or markers a few centimorgans away from the causal gene. However, markers not closely linked or in repulsion phase are highly inconvenient to breeders that aim for marker-assisted selection.

Although our SNP discovery panel was quite comprehensive, some ascertainment bias will remain a problem. The European cultivated gene pool has been enriched with a great diversity of new haplotypes due to introgression breeding. Breeders have used material from the Commonwealth Potato Collection (UK) (Bradshaw and Ramsay 2005), the Dutch-German gene bank (BGRC clones – Braunschweig, Germany) and US plant introductions (PI numbers, Sturgeon Bay, WI). Some of this genetic variation can be traced whether they were included in the discovery panel of 83 varieties, but we certainly have an ascertainment bias with respect to East German and Polish material based on e.g. the Groß Lüsewitzer Kartoffel-Sortimente (GLKS-clones, IPK genebank) and the Vavilov Institute (VIR-clones, St.Petersburg, Russia). Additionally the scientific community is identifying new R-genes (e.g. Bryan et al. 2004; Verzaux et al. 2011; Verzaux et al. 2012). These sources are used by breeders to improve potato varieties, but it is not expected that SolSTW (**Chapter 2**) or SolCAP (Felcher et al. 2012) arrays detect these newly introgressed haplotypes. Furthermore, it is unknown how our SNP array will perform on cultivated material from Latin America. Most likely, material from wild *Solanum* species can't be analysed with the SolSTW SNP array.

Nevertheless, the SolCAP 8303 array has been used on a diverse set of wild relatives of potato (Hardigan et al. 2015). In this study ascertainment bias plays an important role. Analyses of the supplementary files from the study by Hardigan et al. (2015) shows that the highest proportion of heterozygous SNPs is found in the variety Atlantic. This is not unexpected, because Atlantic is one of the varieties in the SNP discovery study. This is probably not because it is the most heterozygous variety, but the result of a severe overestimation of true heterozygosity, because this variety is the least affected by ascertainment bias. Neither is the near complete homozygosity observed for the most distantly related wild *Solanum* species indicative of self-compatibility, but a reflection of near complete assay failure. Hence, the SolCAP SNPs can't reveal the allelic variation in wild species, but neither can our SolSTW array. Still, the proportion and distribution of SNP assay failure is congruent with genetic distance, which allowed Hardigan et al. (2015) to construct dendrograms with a topology that matched their expectations. In

other words the proportion of homozygous SNP data (the amount of ascertainment bias) in a wild species is a measure of genetic distance compared to the material in the SNP discovery study.
In summary, studies using the current SNP arrays will inevitably suffer from some ascertainment bias. Ascertainment bias can be abolished completely by the use of genotyping by sequencing (GBS). Untargeted re-sequencing is preferred over methods with restriction enzymes (Elshire et al. 2011), because restriction site polymorphisms between alleles will give null-alleles and consequently their own haplotype specific ascertainment bias. The continuously decreasing costs of DNA sequencing will increasingly facilitate GBS-based association studies using large variety panels.

Reproducibility

Reproducibility of significant marker-trait associations is an important parameter for the value of a study. In **Chapter 4** we have shown that some QTL regions with significant SNPs can be reproduced, but the individual significant SNPs underlying the QTL regions could not be reproduced. A useful parameter to determine reproducibility of marker-trait associations is the threshold for significance. The higher the threshold, the higher the probability of finding something real, and the higher the reproducibility. In literature a range of association studies in potato have used different thresholds to define a marker as significant. Ranging from $\alpha = 0.05$ (Li et al. 2008; Li et al. 2010; Baldwin et al. 2011; Li et al. 2013; Schönhals et al. 2016) to $\alpha = 0.001$ ($-^{10}\log(p) = 3$) (Urbany et al. 2011; D'hoop et al. 2014) to $-^{10}\log(p) = 4.3$ (Rosyara et al. 2016). In **Chapter 4** we used a threshold of $-^{10}\log(p) = 4$, which is relative conservative as compared to the majority of the studies mentioned above. A common threshold that accommodates for multiple testing is the Bonferroni correction, and that will certainly reduce the number of false positive associations. This correction divides a α of 0.05 by the number of single marker analyses performed. In our study this would result in a significance threshold of $0.05/11674 = -^{10}\log(p)$ of 5.4. This Bonferroni correction is often considered as too strict because SNP assays are not completely independent due to linkage disequilibrium between SNP alleles. In our study a strict Bonferroni correction would result in only two significant SNPs for total glycoalkaloid content (**Chapter 4**). Therefore we relaxed our threshold to $-^{10}\log(p) = 4$ ($\alpha = 0.0001$).
From Manhattan plots the effects of threshold choice are easily understood. Decreasing the threshold to $\alpha = 0.05$, 0.01 or 0.001 will dramatically decrease the number of (false) association. For example 925 SNPs are associated according to $\alpha < 0.05$ for total glycoalkaloid content with kinship correction (**Chapter 4**). Most likely the vast majority of these 925 associations are false positives. When a multiple testing correction is used these associations are indeed assumed to be not significant. However, to my opinion the

significance threshold should not be determined by the available number of markers, because a true association remains a true association independent of the number of markers tested. In other words the threshold of $-^{10}\log(p) = 4$ should be used independent of the number of tested markers. Or such a threshold should be determined based on population size, but also on how significant SNPs are compared to background noise. **Fig. 2** in **Chapter 2** nicely illustrates that indeed a few peaks are clearly different compared to the background noise and all exceeding the threshold of $-^{10}\log(p) = 4$. On the other hand I illustrate in **Fig. 1** of this chapter that a $-^{10}\log(p)$ of 4 is completely irrelevant, because the background noise easily reaches $-^{10}\log(p)$ of 10 or even higher.

Marker density and LD-decay

The probability to detect QTLs in a GWAS can be improved by increasing the number of markers on a SNP array. When a SNP array is developed, it is helpful to know how many markers are required to identify all relevant QTL. The minimal number of markers required for a genome wide association study can be estimated with a genome wide estimation of LD-decay. In **Chapter 3** we describe the patterns of LD and LD-decay in potato. We have shown that there is a large variability in the size of the LD-blocks. For example introgression segments identified in **Chapter 4** and **5** are large while the size of the LD-blocks in young varieties is much smaller. Based on the LD-decay we have proposed that 40,000 SNPs would be optimal for QTL identification. The 14,530 SNP on the SolSTW array is only one third of the optimal 40,000. Nevertheless **Fig. 1** illustrates that the marker density on the SolSTW is sufficient to identify QTLs. For example in **Fig. 1b** the peak markers of the GWAS are 45 kb away from *StCDF* and the peak markers in **Fig 1c** are only 10 kb away from the causal gene for yellow flesh, *StChy2* (Wolters et al. 2010). Nevertheless, the SolCAP array would also be suitable for detection of these two QTLs. For example solcap_snp_c2_50302 is 514 kb away from *StCDF* and reaches a $-^{10}\log(p)$ of 30 and solcap_snp_c2_17552 is 250 kb away from *StChy* and reaches a $-^{10}\log(p)$ of 56. These are monogenic traits and therefore relatively easy to detect. However, for polygenic traits a higher marker density is more important. Additionally the SolCAP SNPs are relatively far away from the causal genes, which make them less suitable for marker-assisted selection.

Within the *StCDF* region the solcap_snp_c2_50302 is the only significant SolCAP SNP. This SolCAP marker is the same as reported by Mosquera et al. (2016). Also they don't find a significant SolCAP marker more close to *StCDF*. In view of their perceived short range LD-decay they proposed another QTL involved in plant maturity, independent of the *StCDF* locus. Based on our findings we conclude that solcap_snp_c2_50302 and the peak markers near *StCDF* are in fact one locus, because there is significant LD between

solcap_snp_c2_50302 and the peak markers near *StCDF* (r^2 = 0.61). The contradicting conclusions could have been avoided with a higher marker density.

Alternative population types for QTL discovery

As described in the preceding paragraphs there are several uncertainties in interpretation of the results in GWAS, which makes interpretations and application of identified SNPs not straightforward. Several factors play a role in the identification of false positives as true positives and not identifying true positives. After recognizing the benefits and limitations of both association panels but also bi-parental populations, two innovative population designs have been proposed to compensate for the limitations of both population types. Multi-parent advanced generation inter-cross (MAGIC) populations were developed in *Arabidopsis thaliana* (Korver et al. 2009) and later also applied in wheat (Huang et al. 2012), rice (Bandillo et al. 2013) and in tomato (Pascual et al. 2015). In maize a similar approach was used and a Nested Association Mapping (NAM) population was developed (McMullen et al. 2009). These innovative population types profit from high number of segregating alleles and increased mapping resolution compared to traditional bi-parental populations, but do not suffer from missing heritability from low frequent alleles and/or population structure. The MAGIC and NAM population types work very well for diploid and/or inbreeding species such as tomato, wheat, rice and maize, however for potato this is not feasible yet because of inbreeding depression. Perhaps in the near future the F_1 hybrid breeding technology (Lindhout et al. 2011) would allow MAGIC and or NAM population development. Meanwhile, an alternative for potato would be a population based on a diallel crossing scheme between 3 to 5 commercially relevant varieties. Such a population has three advantages compared to bi-parental and association populations. **(1)** Similar to an association panel many alleles are segregating, including rare haplotypes, however a more balanced allele frequency allows a better detection of QTLs. **(2)** It is a balanced design such as a bi-parental, NAM or MAGIC population and therefore it does not suffer from false positive and negative associations due to population structure. And **(3)** grouping the three to five half-sib sub-populations, sharing the same parent, will maintain a reasonable mapping resolution of the parental haplotypes.

Towards Marker Assisted Selection in potato

In the previous paragraphs I mainly discuss QTL discovery. In summary GWAS can be important to identify many QTLs, however the experimental design has to be optimized. I made a few suggestions to improve the power of QTL detection with GWAS. Additionally the downstream application of SNP markers in a potato breeding program needs some additional research. I want to address one major risk of the application of SNP markers in marker-assisted selection with the following example that also can explain the low reproducibility of significant SNP markers in **Chapter 4**.

Irreproducibility may arise, when a SNP-allele is not haplotype specific and is present on two haplotypes of which one has a positive effect on a trait and the other no effect on a trait. In an association panel where the haplotype with a positive effect has a higher frequency compared to the haplotype without an effect this SNP is likely to be significantly associated with the trait. Unfortunately, several varieties will have the SNP-allele, but will not have the haplotype that has a positive contribution to the trait. Using such a SNP in marker-assisted selection will not result in the expected results.
Therefore I conclude that knowledge on haplotype specificity of a SNP is an essential piece of information for the successful application of SNP-markers in marker-assisted selection. In the studies of Uitdewilligen (2012) and Wolters et al. (2010) it was shown that several SNPs are unique for a haplotype, whereas other SNPs may lump several haplotypes. Wolters et al. (2010) showed that the causal haplotype for the yellow flesh colour is an example of a haplotype tagged by several hs-SNPs. The haplotype specificity can be an explanation for the extreme low p-value in **Fig 1c**. Unfortunately, haplotype specificity is unknown for almost all SNPs on the SolSTW SNP array. Nevertheless, based on the SNP dating as performed in **Chapter 2** we can assume that a large part of the introgression segments are tagged by haplotype specific SNPs and are therefore suitable for marker assisted selection. For example the SNPs tagging the haplotypes involved in PVY and nematode resistance, presented in **Chapter 5**. For pre-1945 SNPs the underlying haplotypes remain unknown and therefore marker-assisted selection

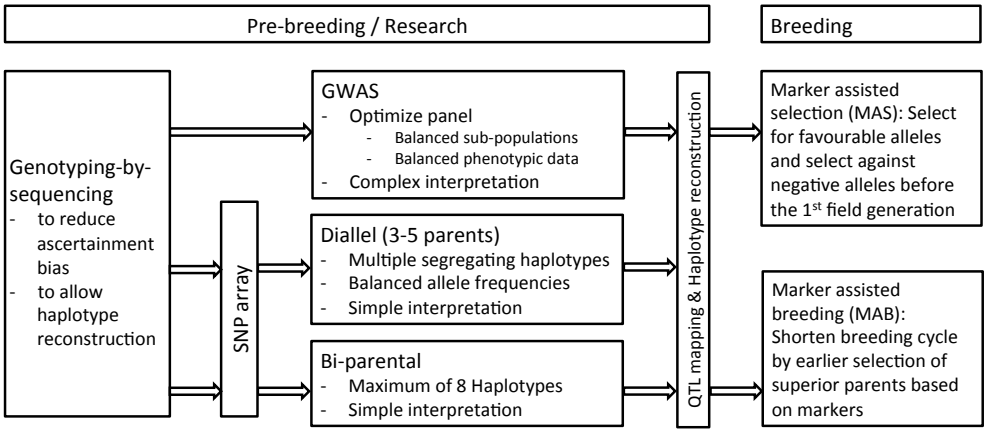

Fig. 2 Schematic overview of potential next steps towards application of marker assisted selection in potato breeding.

QTL discovery can be performed with three population types, each having its own advantages and limitations as described in this chapter. In order to successfully apply the identified QTLs from all three population types, a germplasm wide haplotype reconstruction has to be performed. With a known haplotype structure markers can be applied in marker-assisted breeding (MAB) and marker-assisted selection (MAS). MAS allows breeder to increase population sizes and MAB allows breeders to shorten their breeding cycle, both processes can potentially lead to an increase of genetic gain in potato.

using those SNPs might not result in the expected results. A haplotype map of every gene in the potato genome is essential for successful application of marker-assisted selection.

To sum it all up **(Fig.2**), for QTL discovery I recommend an optimized association panel genotyped with GBS (genotyping by sequencing), a diallel or bi-parental populations genotyped with a SNP array. After QTL detection in all population types a germplasm wide haplotype reconstruction is needed to know which haplotypes are underlying the QTLs. Finally, having sets of markers available for beneficial haplotypes of important traits allows breeders to perform marker-assisted selection/breeding. The markers can be used to select for superior genotypes before the first field generation and they can be used to select superior parents earlier in a breeding program. The former can lead to a larger size of a breeding program and the latter can lead to a shorter breeding cycle. Both processes will increase the genetic gain in potato.

References

Achenbach U, Paulo J, Ilarionova E, Lübeck J, Strahwald J, Tacke E, Hofferbert H-R, Gebhardt C (2008) Using SNP markers to dissect linkage disequilibrium at a major quantitative trait locus for resistance to the potato cyst nematode *Globodera pallida* on potato chromosome V. Theoretical and Applied Genetics 118:619-629

Adetunji I, Willems G, Tschoep H, Bürkholz A, Barnes S, Boer M, Malosetti M, Horemans S, van Eeuwijk F (2014) Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. Theoretical and applied genetics 127:559-571

Anithakumari AM, Tang J, van Eck HJ, Visser RGF, Leunissen JAM, Vosman B, van der Linden CG (2010) A pipeline for high throughput detection and mapping of SNPs from EST databases. Molecular breeding 26:65-75

Anonymous (2015) Cultuur en gebruikswaardeonderzoek aardappelen (CGO-aardappelen). https://www.plantum.nl/321519683/Publicatie?newsitemid=130613250

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627-631

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, 3(10):e3376.

Bakker E, Borm T, Prins P, van der Vossen E, Uenk G, Arens M, de Boer J, van Eck HJ, Muskens M, Vossen J, van der Linden G, van Ham R, Klein-Lankhorst R, Visser RGF, Smant G, Bakker J, Goverse A (2011) A genome-wide genetic map of NB-LRR disease resistance loci in potato. Theoretical and Applied Genetics 123:493-508

Baldwin SJ, Dodds KG, Auvray B, Genet RA, Macknight RC, Jacobs JME (2011) Association mapping of cold-induced sweetening in potato using historical phenotypic data. Annals of Applied Biology 158:248-256

Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT, Dilla-Ermita CJ, Tung C-W, McCouch S, Thomson M, Mauleon R, Singh RK, Gregorio G, Redoña E, Leung H (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6:1-15

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. The Plant Journal 51:910-918

Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana* 11:867-879

Bonierbale MW, Plaisted RL, Tanksley SD (1988) RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics 120:1095-1103

Bourke PM, Voorrips RE, Visser RG, Maliepaard C (2015) The Double-Reduction Landscape in Tetraploid Potato as Revealed by a High-Density Linkage Map. Genetics 201:853-863

Bradshaw EJ, Dale BMF, Mackay RG (2003) Use of mid-parent values and progeny tests to increase the efficiency of potato breeding for combined processing quality and disease and pest resistance. Theoretical and Applied Genetics 107:36-42

Bradshaw J, Ramsay G (2005) Utilisation of the Commonwealth Potato Collection in potato breeding. Euphytica 146(1-2):9-19.

Bradshaw JE (2009) Potato breeding at the Scottish plant breeding station and the Scottish Crop Research Institute: 1920–2008. Potato research 52:141-172

Bradshaw JE, Hackett CA, Pande B, Waugh R, Bryan GJ (2008) QTL mapping of yield, agronomic and quality traits in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*). Theoretical and Applied Genetics 116:193-211

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, Ben C, Denny R, Sadowsky MJ, Ronfort J, Bataillon T, Young ND, Tiffin P (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proceedings of the National Academy of Sciences 108:E864–E870

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-1177

Brigneti G, Garcia-Mas J, Baulcombe DC (1997) Molecular mapping of the potato virus Y resistance gene Rysto in potato. Theoretical and Applied Genetics 94:198-203

Bryan G, McLean K, Bradshaw J, De Jong W, Phillips 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

Bryan GJ, McLean K, Pande B, Purvis A, Hackett CA, Bradshaw JE, Waugh R (2004) Genetical dissection of H3-mediated polygenic PCN resistance in a heterozygous autotetraploid potato population. Molecular Breeding 14:105-116

Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnology Journal 7:347-354

Cárdenas PD, Sonawane PD, Heinig U, Bocobza SE, Burdman S, Aharoni A (2015) The bitter side of the nightshades: Genomics drives discovery in *Solanaceae* steroidal alkaloid metabolism. Phytochemistry 113:24-32

Cardenas PD, Sonawane PD, Pollier J, Vanden Bossche R, Dewangan V, Weithorn E, Tal L, Meir S, Rogachev I, Malitsky S, Giri AP, Goossens A, Burdman S, Aharoni A (2016) GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. Nat Commun 7

Carputo D, Terra A, Barone A, Esposito F, Fogliano V, Monti L, Frusciante L (2003) Glycoalkaloids and acclimation capacity of hybrids between *Solanum tuberosum* and the incongruent hardy species *Solanum commersonii*. Theoretical and Applied Genetics 107:1187-1194

Celebi-Toprak F, Slack AS, Jahn MM (2002) A new gene, Ny tbr , for hypersensitivity to Potato virus Y from *Solanum tuberosum* Maps to Chromosome IV. Theoretical and Applied Genetics 104:669-674

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WTB, Ramsay L, Mackay I, Balding DJ, Consortium TA, Waugh R, O'Sullivan DM (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proceedings of the National Academy of Sciences 107:21611-21616

Comadran J, Ramsay L, MacKenzie K, Hayes P, Close TJ, Muehlbauer G, Stein N, Waugh R (2011) Patterns of polymorphism and linkage disequilibrium in cultivated barley. Theoretical and applied genetics 122:523-531

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

D'hoop BB (2009) Association mapping in tetraploid potato. Wageningen Universiteit (Wageningen University)

D'hoop BB, 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 121:1151-1170

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 54:157-181

D'hoop BB, Keizer PLC, 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 127:731-748

De Jong H, Proudfoot KG, Murphy AM (2001) The germplasm release of F87084, a fertile, adapted clone with multiple disease resistances. American Journal of Potato Research 78:141-149

Dellaert LM, Vinke HJ (1987) Testing potatoes for resistance to Globodera pallida pathotype Pa-3; resistance spectra of plant genotypes and virulence spectra of Pa-3 isolates. Revue Nématol. 10(4):445-453.

Delourme R, Falentin C, Fomeju BF, Boillot M, Lassalle G, André I, Duarte J, Gauthier V, Lucante N, Marty A, Pauchon M, Pichon J-P, Ribière N, Trotoux G, Blanchard P, Rivière N, Martinant J-P, Pauquet J (2013) High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. BMC Genomics 14:1-18

Douches D, Maas D, Jastrzebski K, Chase R (1996) Assessment of potato breeding progress in the USA over the last century. Crop Science 36:1544-1552

El-Kharbotly A, Leonards-Schippers C, Huigen DJ, Jacobsen E, Pereira A, Stiekema WJ, Salamini F, Gebhardt C (1994) Segregation analysis and RFLP mapping of the R1 and R3 alleles conferring race-specific resistance to *Phytophthora infestans* in progeny of dihaploid potato parents. Molecular and General Genetics MGG 242:749-754

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6:e19379

Endelman J, Jansky S (2016) Genetic mapping with an inbred line derived F2 population in potato. Theoretical and Applied Genetics In press

Esteras C, Formisano G, Roig C, Díaz A, Blanca J, Garcia-Mas J, Gómez-Guillamón ML, López-Sesé AI, Lázaro A, Monforte AJ, Picó B (2013) SNP genotyping in melons: genetic variation, population structure, and linkage disequilibrium. Theoretical and Applied Genetics 126:1285-1303

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

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

Fischer M, Schreiber L, Colby T, Kuckenberg M, Tacke E, Hofferbert H-R, Schmidt J, Gebhardt C (2013) Novel candidate genes influencing natural variation in potato tuber cold sweetening identified by comparative proteomics and association mapping. BMC plant biology 13:1

Finkers-Tomczak A, Danan S, van Dijk T, Beyene A, Bouwman L, Overmars H, van Eck HJ, Goverse A, Bakker J, Bakker E (2009) A high-resolution map of the Grp1 locus on chromosome V of potato harbouring broad-spectrum resistance to the cyst nematode species G*lobodera pallida* and *Globodera rostochiensis*. Theoretical and Applied Genetics 119:165-173

Flint-Garcia SA, Thornsberry JM, IV B (2003) Structure of linkage disequilibrium in plants*. Annual review of plant biology 54:357-374

Flis B, Hennig J, Strzelczyk-Żyta D, Gebhardt C, Marczewski W (2005) The Ry-f sto gene from Solanum stoloniferum for extreme resistant to Potato virus Y maps to potato chromosome XII and is diagnosed by PCR marker GP122718 in PVY resistant potato cultivars. Molecular Breeding 15:95-101

Friedman M (2006) Potato glycoalkaloids and metabolites: roles in the plant and in the diet. Journal of Agricultural and Food Chemistry 54:8655-8681

Gebhardt C, Mugniery D, Ritter E, Salamini F, Bonnel E (1993) Identification of RFLP markers closely linked to the H1 gene conferring resistance to *Globodera rostochiensis* in potato. Theoretical and Applied Genetics 85:541-544

Ghislain M, Spooner DM, Rodríguez F, Villamón F, Nunez J, Vásquez C, Waugh R, Bonierbale M (2004) Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. Theoretical and Applied Genetics 108:881-890

Ginzberg I, Thippeswamy M, Fogelman E, Demirel U, Mweetwa AM, Tokuhisa J, Veilleux RE (2012) Induction of potato steroidal glycoalkaloid biosynthetic pathway by overexpression of cDNA encoding primary metabolism HMG-CoA reductase and squalene synthase. Planta 235:1341-1353

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:e63939

Hämäläinen HJ, Watanabe NK, Valkonen TJP, Arihara A, Plaisted LR, Pehu E, Miller L, Slack AS (1997) Mapping and marker-assisted selection for a gene for extreme resistance to potato virus Y. Theoretical and Applied Genetics 94:192-197

Hämäläinen HJ, Sorri AV, Watanabe NK, Gebhardt C, Valkonen TJP (1998) Molecular examination of a chromosome region that controls resistance to potato Y and A potyviruses in potato. Theoretical and Applied Genetics 96:1036-1043

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 12:1

Hamilton JP, Sim S-C, Stoffel K, Van Deynze A, Buell CR, Francis DM (2012) Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. The Plant Genome 5:17-29

Hardigan MA, Bamberg J, Buell CR, Douches DS (2015) Taxonomy and Genetic Differentiation among Wild and Cultivated Germplasm of sect. The Plant Genome 8

Hawkes JG, Francisco-Ortega J (1993) The early history of the potato in Europe. Euphytica 70:1-7

Hehl R, Faurie E, Hesselbach J, Salamini F, Whitham S, Baker B, Gebhardt C (1999) TMV resistance gene N homologues are linked to *Synchytrium endobioticum* resistance in potato. Theoretical and Applied Genetics 98:379-386

Heldák J, Bežo M, Štefúnová V, Galliková A (2007) Selection of DNA Markers for Detection of Extreme Resistance to Potato virus Y in Tetraploid Potato (*Solanum tuberosum* L.) F1 Progenies. Czech J. Genet. Plant Breed. 43(4):125-134.

Hellenäs KE, Branzell C, Johnsson H, Slanina P (1995) High levels of glycoalkaloids in the established Swedish potato variety Magnum Bonum. Journal of the Science of Food and Agriculture 68:249-255

Higashide T, Heuvelink E (2009) Physiological and morphological changes over the past 50 years in yield components in tomato. Journal of the American Society for Horticultural Science 134:460-465

Hirsch CN, Hirsch CD, Felcher K, Coombs J, Zarka D, Van Deynze A, De Jong W, Veilleux RE, Jansky S, Bethke P, Douches DS, Buell CR (2013) Retrospective View of North American Potato (*Solanum tuberosum* L.) Breeding in the 20th and 21st Centuries. G3: Genes|Genomes|Genetics 3:1003-1013

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6:95-108

Hoagland RE (2009) Toxicity of tomatine and tomatidine on weeds, crops and phytopathogenetic fungi. Allelopathy J 23:425-436

Hoekstra AY, Chapagain AK (2011) Globalization of water: Sharing the planet's freshwater resources. John Wiley & Sons

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant biotechnology journal 10:826-839

Hutvágner G, Bánfalvi Z, Milánkovics I, Silhavy D, Polgár Z, Horváth S, Wolters P, Nap J-P (2001) Molecular markers associated with leptinine production are located on chromosome *1* in *Solanum chacoense*. Theoretical and Applied Genetics 102:1065-1071

Hyten DL, Choi I-Y, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175:1937-1944

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, Tikunov Y, Bovy A, Chikate Y, Singh P, Rogachev I, Beekwilder J, Giri AP, Aharoni A (2013) Biosynthesis of Antinutritional Alkaloids in Solanaceous Crops Is Mediated by Clustered Genes. Science 341:175-179

Jupe F, Pritchard L, Etherington GJ, MacKenzie K, Cock PJ, Wright F, Sharma SK, Bolser D, Bryan GJ, Jones JD, Hein I (2012) Identification and localisation of the NB-LRR gene family within the potato genome. BMC Genomics 13:1-14

Khan MS (2012) Assessing genetic variation in growth and development of potato. publisher not identified

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nature Genetics 39:1151-1155

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nature genetics 39:1151-1155

Kloosterman B, Oortwijn M, America T, de Vos R, Visser RGF, Bachem CWB (2010) From QTL to candidate gene: genetical genomics of simple and complex traits in potato using a pooling strategy. BMC genomics 11:158

Kloosterman B, Abelenda JA, Gomez MDMC, Oortwijn M, de Boer JM, Kowitwanich K, Horvath BM, Van Eck HJ, Smaczniak C, Prat S, Visser RGF, Bachem, CW (2013). Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495(7440):246-250.

Kooke R, Kruijer W, Bours R, Becker FFM, Kuhn A, Buntjer J, Doeswijk T, Guerra J, Bouwmeester HJ, Vreugdenhil D, Keurentjes JJB (2016) Genome-wide association mapping and genomic prediction elucidate the genetic architecture of morphological traits in *Arabidopsis thaliana*. Plant physiology:pp. 00997.02015

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

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. PLoS Genet 5:e1000551

Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435-446

Krits P, Fogelman E, Ginzberg I (2007) Potato steroidal glycoalkaloid levels and the expression of key isoprenoid metabolic genes. Planta 227:143-150

Laidig F, Piepho H-P, Drobek T, Meyer U (2014) Genetic and non-genetic long-term trends of 12 different crops in German official variety performance trials and on-farm yield trends. Theoretical and Applied Genetics 127:2599-2617

Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS-M, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053-1059

Li L, Paulo M-J, Strahwald J, Lübeck J, Hofferbert H-R, Tacke E, Junghans H, Wunder J, Draffehn A, van Eeuwijk F, Gebhardt C (2008) Natural DNA variation at candidate loci is associated with potato chip color, tuber starch content, yield and starch yield. Theoretical and Applied Genetics 116:1167-1181

Li L, Paulo MJ, van Eeuwijk FA, Gebhardt C (2010) Statistical epistasis between candidate gene alleles for complex tuber traits in an association mapping population of tetraploid potato. Theoretical and applied genetics 121:1303-1310

Li L, Tacke E, Hofferbert H-R, Lübeck J, Strahwald J, Draffehn AM, Walkemeier B, Gebhardt C (2013) Validation of candidate gene markers for marker-assisted selection of potato cultivars with improved tuber quality. Theoretical and Applied Genetics 126:1039-1052

Li X, Van Eck HJ, Rouppe van der Voort JNAM, 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, Han Y, Wei Y, Acharya A, Farmer AD, Ho J, Monteros MJ, Brummer EC (2014) Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. PLoS One 9:e84329

Lindhout P, Meijer D, Schotte T, Hutten RCB, Visser RGF, van Eck HJ (2011) Towards F1 hybrid seed potato breeding. Potato Research 54:301-312

Lindqvist-Kreuze H, Gastelo M, Perez W, Forbes GA, de Koeyer D, Bonierbale M (2014) Phenotypic Stability and Genome-Wide Association Study of Late Blight Resistance in Potato Genotypes Adapted to the Tropical Highlands. Phytopathology 104:624-633

Long NV, Dolstra O, Malosetti M, Kilian B, Graner A, Visser RGF, van der Linden CG (2013) Association mapping of salt tolerance in barley (*Hordeum vulgare* L.). Theoretical and applied genetics 126:2335-2351

Lopez-Pardo R, Barandalla L, Ritter E, Ruiz de Galarreta JI (2013) Validation of molecular markers for pathogen resistance in potato. Plant Breed 132:246-251

Love SL, Pavek JJ, Thompson-Johns A, Bohl W (1998) Breeding progress for potato chip quality in North American varieties. American Journal of Potato Research 75:27-36

Maccaferri M, Sanguineti MC, Noli E, Tuberosa R (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Molecular Breeding 15:271-290

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 175:879-889

Manrique-Carpintero NC, Tokuhisa JG, Ginzberg I, Holliday JA, Veilleux RE (2013) Sequence diversity in coding regions of candidate genes in the glycoalkaloid biosynthetic pathway of wild potato species. G3: Genes| Genomes| Genetics 3:1467-1479

Manrique-Carpintero NC, Tokuhisa JG, Ginzberg I, Veilleux RE (2014) Allelic variation in genes contributing to glycoalkaloid biosynthesis in a diploid interspecific population of potato. Theoretical and applied genetics 127:391-405

Mariot RF, de Oliveira LA, Voorhuijzen M, Staats M, Hutten R, van Dijk JP, Kok E, Frazzon J (2016) Characterization and transcriptional profile of genes involved on glycoalkaloid biosynthesis in new varieties of *Solanum tuberosum* L. Journal of Agricultural and Food Chemistry

McCord PH, Sosinski BR, Haynes KG, Clough ME, Yencho GC (2011) Linkage Mapping and QTL Analysis of Agronomic Traits in Tetraploid Potato (subsp.). Crop science 51:771-785

McCue KF, Rockhold DR, Chhan A, Belknap WR (2011) Structure of two *Solanum tuberosum* steroidal glycoalkaloid glycosyltransferase genes and expression of their promoters in transgenic potatoes. American journal of potato research 88:485-492

McCue KF, Shepherd LV, Allen PV, Maccree MM, Rockhold DR, Corsini DL, Davies HV, Belknap WR (2005) Metabolic compensation of steroidal glycoalkaloid biosynthesis in transgenic potato tubers: using reverse genetics to confirm the in vivo enzyme function of a steroidal alkaloid galactosyltransferase. Plant Science 168:267-273

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Rosas MO, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) Genetic Properties of the Maize Nested Association Mapping Population. Science 325:737-740

Medina T, Fogelman E, Chani E, Miller A, Levin I, Levy D, Veilleux R (2002) Identification of molecular markers associated with leptine in reciprocal backcross families of diploid potato. Theoretical and Applied Genetics 105:1010-1018

Menéndez CM, Ritter E, Schäfer-Pregl R, Walkemeier B, Kalde A, Salamini F, Gebhardt C (2002) Cold sweetening in diploid potato: mapping quantitative trait loci and candidate genes. Genetics 162:1423-1434

Milbourne D, Meyer R, Collins A, Ramsay L, Gebhardt C, Waugh R (1998) Isolation, characterisation and mapping of simple sequence repeat loci in potato. Molecular and General Genetics MGG 259:233-245

Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WTB, Flavell AJ, Marshall D (2010) Flapjack—graphical genotype visualization. Bioinformatics 26:3133-3134

Moehs CP, Allen PV, Friedman M, Belknap WR (1997) Cloning and expression of solanidine UDP-glucose glucosyltransferase from potato. The Plant Journal 11:227-236

Moragues M, Comadran J, Waugh R, Milne I, Flavell A, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theoretical and Applied Genetics 120:1525-1534

Mosquera T, Alvarez MF, Jiménez-Gómez JM, Muktar MS, Paulo MJ, Steinemann S, Li J, Draffehn A, Hofmann A, Lübeck J, Strahwald J, Tacke E, Hofferbert H-R, Walkemeier B, Gebhardt C (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:e0156254

Moury B, Caromel B, Johansen E, Simon V, Chauvin L, Jacquot E, Kerlan C, Lefebvre V (2011) The Helper Component Proteinase Cistron of Potato virus Y Induces Hypersensitivity and Resistance in Potato Genotypes Carrying Dominant Resistance Genes on Chromosome IV. Molecular Plant-Microbe Interactions 24:787-797

Myles S, Chia J-M, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D (2010) Rapid genomic characterization of the genus vitis. PloS one 5:e8219

Nie X, Liang Z, Nie B, Murphy A, Singh M (2015) Studies on varietal response to different strains of Potato virus Y (PVY) reveal hypersensitive resistance in Exploits to PVYO and extreme resistance in F87084 to all tested strains. American Journal of Potato Research 92:23-31

Nenaah G (2011) Individual and synergistic toxicity of solanaceous glycoalkaloids against two coleopteran stored-product insects. Journal of pest science 84:77-86

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190-193

Obidiegwu JE, Sanetomo R, Flath K, Tacke E, Hofferbert H-R, Hofmann A, Walkemeier B, Gebhardt C (2015) Genomic architecture of potato resistance to *Synchytrium endobioticum* disentangled using SSR markers and the 8.3 k SolCAP SNP genotyping array. BMC genetics 16:1

Paal J, Henselewski H, Muth J, Meksem K, Menéndez CM, Salamini F, Ballvora A and Gebhardt C (2004), Molecular cloning of the potato *Gro1-4* gene conferring resistance to pathotype Ro1 of the root cyst nematode *Globodera rostochiensis*, based on a candidate gene approach. The Plant Journal 38(2):285-297. doi: 10.1111/j.1365-313X.2004.02047.x

Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet JP, Le QH, Chauchard B, Verschave P, Causse M (2015) Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. Plant biotechnology journal 13:565-577

PGSC (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189-195

Prashar A, Hornyik C, Young V, McLean K, Sharma S, Dale MF, Bryan G (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 127:2159-2171.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Raboin L-M, Pauquet J, Butterfield M, D'Hont A, Glaszmann J-C (2008) Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. Theoretical and Applied Genetics 116:701-714

Ray DK, Mueller ND, West PC, Foley JA (2013) Yield trends are insufficient to double global crop production by 2050. PloS one 8:e66428

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences 98:11479-11484

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proceedings of the National Academy of Sciences 109:8872-8877

Rijk B, van Ittersum M, Withagen J (2013) Genetic progress in Dutch crop yields. Field Crops Research 149:262-268

Ronning CM, Stommel JR, Kowalski SP, Sanford LL, Kobayashi RS, Pineada O (1999) Identification of molecular markers associated with leptine production in a population of *Solanum chacoense* Bitter. Theoretical and applied genetics 98:39-46

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

Rouppe van der Voort J, Kanyuka K, van der Vossen E, Bendahmane A, Mooijman P, Klein-Lankhorst R, Stiekema W, Baulcombe D, Bakker J (1999) Tight physical linkage of the nematode resistance gene *Gpa2* and the virus resistance gene Rx on a single segment introgressed from the wild species Solanum tuberosum subsp. andigena CPC 1673 into cultivated potato. Molecular Plant Microbe Interactions 12:197-206.

Rouppe van der Voort JNAM, Van der Vossen E, Bakker E, Overmars H, Van Zandvoort P, Hutten R, Klein Lankhorst R, Bakker J (2000) Two additive QTLs conferring broad-spectrum resistance in potato to *Globodera pallida* are localized on resistance gene clusters. Theoretical and applied genetics 101:1122-1130

Sagredo B, Lafta A, Casper H, Lorenzen J (2006) Mapping of genes associated with leptine content of tetraploid potato. Theoretical and Applied Genetics 114:131-142

Sagredo B, Lorenzen J, Casper H, Lafta A (2011) Linkage analysis of a rare alkaloid present in a tetraploid potato with *Solanum chacoense* background. Theoretical and applied genetics 122:471-478

Salaman RN (1985) The history and social influence of the potato. Cambridge University Press

Sato M, Nishikawa K, Komura K, Hosaka K (2006) Potato Virus Y Resistance Gene, Ry chc, Mapped to the Distal End of Potato Chromosome 9. Euphytica 149:367-372

Sattarzadeh A, Achenbach U, Lübeck J, Strahwald J, Tacke E, Hofferbert H-R, Rothsteyn T, Gebhardt C (2006) Single nucleotide polymorphism (SNP) genotyping as basis for developing a PCR-based marker highly diagnostic for potato varieties with high resistance to *Globodera pallida* pathotype Pa2/3. Molecular Breeding 18:301-312

Sawai S, Ohyama K, Yasumoto S, Seki H, Sakuma T, Yamamoto T, Takebayashi Y, Kojima M, Sakakibara H, Aoki T, Muranaka T, Saito K, Umemoto N (2014) Sterol Side Chain Reductase 2 Is a Key Enzyme in the Biosynthesis of Cholesterol, the Common Precursor of Toxic Steroidal Glycoalkaloids in Potato. The Plant Cell 26:3763-3774

Schäfer-Pregl R, Ritter E, Hesselbach J, Lovatti L, Walkemeier B, Thelen H, Salamini F, Gebhardt C (1998) Analysis of quantitative trait loci (QTLs) and quantitative trait alleles (QTAs) for potato tuber yield and starch content. Theoretical and Applied Genetics 97:834-846

Schönhals EM, Ortega F, Barandalla L, Aragones A, Ruiz de Galarreta JI, Liao J-C, Sanetomo R, Walkemeier B, Tacke E, Ritter E, 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 129:767-785

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: Genes| Genomes| Genetics 4:1797-1811

Sharma R, Bhardwaj V, Dalamu D, Kaushik SK, Singh BP, Sharma S, Umamaheshwari R, Baswaraj R, Kumar V, Gebhardt C (2014) Identification of elite potato genotypes possessing multiple disease resistance genes through molecular approaches. Scientia Horticulturae 179:204-211

Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W, Carboni MF, D'Ambrosio JM, de la Cruz G, Di Genova A, Douches DS, Eguiluz M, Guo X, Guzman F, Hackett CA, Hamilton JP, Li G, Li Y, Lozano R, Maass A, Marshall D, Martinez D, McLean K, Mejía N, Milne L, Munive S, Nagy I, Ponce O, Ramirez M, Simon R, Thomson SJ, Torres Y, Waugh R, Zhang Z, Huang S, Visser RGF, Bachem CWB, Sagredo B, Feingold SE, Orjeda G, Veilleux RE, Bonierbale M, Jacobs JME, Milbourne D, Martin DMA, Bryan GJ (2013) Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps. G3: Genes|Genomes|Genetics 3:2031-2047

Sim S-C, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A, Hamilton JP, Buell CR, Causse M, Wijeratne S, Francis DM (2012) Development of a Large SNP Genotyping Array and Generation of High-Density Genetic Maps in Tomato. PLoS ONE 7:e40563

Simko I, Costanzo S, Haynes KG, Christ BJ, Jones RW (2004) Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. Theoretical and Applied Genetics 108:217-224

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

Slater AT, Wilson GM, Cogan NO, Forster JW, Hayes BJ (2014b) Improving the analysis of low heritability complex traits for enhanced genetic gain in potato. Theoretical and applied genetics 127:809-820

Song Y-S, Hepting L, Schweizer G, Hartl L, Wenzel G, Schwarzfischer A (2005) Mapping of extreme resistance to PVY (Ry sto) on chromosome XII using anther-culture-derived primary dihaploid potato lines. Theoretical and applied genetics 111:879-887

Song Y-S, Schwarzfischer A (2008) Development of STS Markers for Selection of Extreme Resistance (Ry sto) to PVY and Maternal Pedigree Analysis of Extremely Resistant Cultivars. American Journal of Potato Research

Sørensen KK, Kirk HG, Olsson K, Labouriau R, Christiansen J (2008) A major QTL and an SSR marker associated with glycoalkaloid content in potato tubers from *Solanum tuberosum* × S. *sparsipilum* located on chromosome I. Theoretical and Applied Genetics 117:1-9

Stich B, Urbany C, Hoffmann P, Gebhardt C (2013) Population structure and linkage disequilibrium in diploid and tetraploid potato revealed by genome-wide high-density genotyping using the SolCAP SNP array. Plant Breeding 132:718-724

Szajko K, Chrzanowska M, Witek K, Strzelczyk-Żyta D, Zagórska H, Gebhardt C, Hennig J, Marczewski W (2008) The novel gene Ny-1 on potato chromosome IX confers hypersensitive resistance to Potato virus Y and is an alternative to Ry genes in potato breeding for PVY resistance. Theoretical and Applied Genetics 116:297-303

Szajko K, Strzelczyk-Żyta D, Marczewski W (2014) Ny-1 and Ny-2 genes conferring hypersensitive response to potato virus Y (PVY) in cultivated potatoes: mapping and marker-assisted selection validation for PVY resistance in potato breeding. Molecular Breeding 34:267-271

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. mays L.). Proceedings of the National Academy of Sciences 98:9161-9166

Thomson M, Zhao K, Wright M, McNally K, Rey J, Tung C-W, Reynolds A, Scheffler B, Eizenga G, McClung A, Kim H, Ismail A, de Ocampo M, Mojica C, Reveche MY, Dilla-Ermita C, Mauleon R, Leung H, Bustamante C, McCouch S (2012) High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. Molecular Breeding 29(4):875-886

Tomczyńska I, Jupe F, Hein I, Marczewski W, Śliwka J (2014) Hypersensitive response to Potato virus Y in potato cultivar Sárpo Mira is conferred by the Ny-Smira gene located on the long arm of chromosome IX. Molecular Breeding 34:471-480

Toxopeus HJ (1956) Reflections on the origin of new physiologic races in Phytophthora infestans and the breeding for resistance in potatoes. Euphytica 5:221-237

Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. Plant Biotechnology Journal 7:334-346

Uitdewilligen J (2012) Discovery and genotyping of existing and induced DNA sequence variation in potato. Wageningen University

Uitdewilligen JGAML, Wolters A-MA, Bjorn B, 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:e62355

United Nations; Department of Economic and Social Affairs PD (2015) World Population Prospects: The 2015 Revision, Key Findings and Advance Tables. ESA/P/WP.241.

Urbany C, Stich B, Schmidt L, Simon L, Berding H, Junghans H, Niehoff K-H, Braun A, Tacke E, Hofferbert H-R, Lübeck J, Strahwald J, Gebhardt C (2011) Association

genetics in Solanum tuberosum provides new insights into potato tuber bruising and enzymatic tissue discoloration. BMC Genomics 12:1-14

Valcarcel J, Reilly K, Gaffney M, O'Brien N (2014) Effect of genotype and environment on the glycoalkaloid content of rare, heritage, and commercial potato varieties. Journal of food science 79:T1039-T1048

Valkonen, JPT (2007) Chapter 28 - Viruses: Economical losses and biotechnological potential, In Potato Biology and Biotechnology, Eds. Vreugdenhil, Bradshaw, Gebhardt, Govers, Mackerron, Taylor and Ross, Elsevier Science B.V., Amsterdam. pp. 619-641 doi:10.1016/B978-044451018-1/50070-1

van Berloo R (1999) Computer note. GGT: software for the display of graphical genotypes. Journal of Heredity 90:328-329

Van Berloo R (2008) GGT 2.0: Versatile software for visualization and analysis of genetic data. J Heredity 99(2):232-236.

Van Berloo R, Hutten R, Van Eck H, Visser R (2007) An online potato pedigree database resource. Potato research 50:45-57

van de Wouw M, van Hintum T, Kik C, van Treuren R, Visser B (2010) Genetic diversity trends in twentieth century crop varieties: a meta analysis. Theoretical and Applied Genetics 120:1241-1252

van Eck HJ, Jacobs JME, van Dijk J, Stiekema WJ, Jacobsen E (1993) Identification and mapping of three flower colour loci of potato (*S. tuberosum* L.) by RFLP analysis. Theoretical and Applied Genetics 86:295-300

van Eck HJ, Jacobs JME, van ven Berg PMMM, Stiekema WJ, Jacobsen E (1994) The inheritance of anthocyanin pigmentation in potato (*Solanum tuberosum* L.) and mapping of tuber skin colour loci using RFLPs. Heredity 73:410-421

van Eck, HJ, Willemsen J, Witteveen A, Visser RGF, Vos PG, Borm TJA (2014) Fine mapping of the *Ro*-locus involved in tuber shape on potato chromosome *10*. Abstract Book, EAPR conference July 2014, Brussels.

Van Inghelandt D, Reif JC, Dhillon BS, Flament P, Melchinger AE (2011) Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. Theoretical and applied genetics 123:11-20

Van Ooijen JW (2004) MapQTL® 5. Software for the mapping of quantitative trait loci in experimental populations Kyazma BV, Wageningen

van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan GJ, Caromel B, Ghareeb B, Isidore E, de Jong W, van Koert P, Lefebvre V, Milbourne D, Ritter E, Rouppe van der Voort JNAM, Rousselle-Bourgeois F, van Vliet J, Waugh R, Visser RGF, Bakker J, van Eck HJ (2006) Construction of a 10,000-Marker Ultradense

Genetic Recombination Map of Potato: Providing a Framework for Accelerated Gene Isolation and a Genomewide Physical Map. Genetics 173:1075-1087

Verzaux E, Budding D, de Vetten N, Niks RE, Vleeshouwers VG, van der Vossen EA, Jacobsen E, Visser RGF (2011) High resolution mapping of a novel late blight resistance gene Rpi-avl1, from the wild Bolivian species *Solanum avilesii*. American journal of potato research 88:511-519

Verzaux E, van Arkel G, Vleeshouwers VG, van der Vossen EA, Niks RE, Jacobsen E, Vossen J, Visser RGF (2012) High-resolution mapping of two broad-spectrum late blight resistance genes from two wild species of the *Solanum circaeifolium* group. Potato Research 55:109-123

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

Voorrips RE, Maliepaard CA (2012) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC bioinformatics 13:1

Vos P, Hogers R, Bleeker M, Reijans M, Van de Lee T, Hornes M, Friters A, Pot J, Paleman J, Kuiper M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic acids research 23:4407-4414

Vos PG, Uitdewilligen JGAML, 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 128:2387-2401

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., International Wheat Genome Sequencing Consortium, Lillemo, M., Mather, D., Appels, R., Dolferus, R., Brown-Guedira, G., Korol, A., Akhunova, A. R., Feuillet, C., Salse, J., Morgante, M., Pozniak, C., Luo, M.-C., Dvorak, J., Morell, M., Dubcovsky, J., Ganal, M., Tuberosa, R., Lawley, C., Mikoulitch, I., Cavanagh, C., Edwards, K. J., Hayden, M. and Akhunov, E. (2014), Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant Biotechnol J, 12: 787–796. doi:10.1111/pbi.12183

Wang Y-H, Upadhyaya HD, Burrell AM, Sahraeian SME, Klein RR, Klein PE (2013) Genetic structure and linkage disequilibrium in a diverse, representative collection of the C4 model plant, Sorghum bicolor. G3: Genes| Genomes| Genetics 3:783-793

Werij JS, Kloosterman B, Celis-Gamboa C, de Vos CHR, America T, Visser RGF, Bachem CWB (2007) Unravelling enzymatic discoloration in potato through a

combined approach of candidate genes, QTL, and expression analysis. Theoretical and Applied Genetics 115:245-252

Wolters A-MA, Uitdewilligen JGAML, 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 73:659-671

Würschum T, Langer SM, Longin CFH, Korzun V, Akhunov E, Ebmeyer E, Schachschneider R, Schacht J, Kazman E, Reif JC (2013) Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. Theoretical and Applied Genetics 126:1477-1486

Yamamoto T, Nagasaki H, Yonemaru J-i, Ebana K, Nakajima M, Shibaya T, Yano M (2010) Fine definition of the pedigree haplotypes of closely related rice varieties by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics 11(1):267

Yan J, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS one 4:e8451

Yencho GC, Kowalski SP, Kobayashi RS, Sinden SL, Bonierbale MW, Deahl KL (1998) QTL mapping of foliar glycoalkaloid aglycones in *Solanum tuberosum* × *S. berthaultii* potato progenies: quantitative variation and plant secondary metabolism. Theoretical and applied genetics 97:563-574

Young ND, Tanksley SD (1989) Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theoretical and Applied Genetics 77:95-101

Yun SJ, Gyenis L, Bossolini E, Hayes PM, Matus I, Smith KP, Steffenson BJ, Tuberosa R, Muehlbauer GJ (2006) Validation of Quantitative Trait Loci for Multiple Disease Resistance in Barley Using Advanced Backcross Lines Developed with a Wild Barley Both S.J. Yun and L. Gyenis contributed equally to this work. Crop Science 46:1179-1186

Zegeye H, Rasheed A, Makdis F, Badebo A, Ogbonnaya FC (2014) Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. PloS one 9:e105593

Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun 2:467

Zitnak A, Johnston GR (1970) Glycoalkaloid content of B5141-6 potatoes. American Potato Journal 47:256-260

Summary

In this thesis the results are described of investigations of various application of genome wide SNP (single nucleotide polymorphism) markers. The set of SNP markers was identified by GBS (genotyping by sequencing) strategy. The resulting dataset of 129,156 SNPs across 83 tetraploid varieties was used directly to map traits, but also as a basis for the development of a 20K SNP array in Potato (*Solanum tuberosum* L.). Subsequently this array, named SolSTW, was used to collect genotypic data from 569 potato genotypes. This dataset offered insight in the breeding history of potato, population structure, linkage disequilibrium (LD) and the potential of GWAS (genome wide association studies) in potato.

In **Chapter 2** we describe to development of the SolSTW 20K Infinium SNP array. One third of the SNPs on this array originate from the well-known SolCAP 8303 SNP array. The other SNPs are a subset from a targeted re-sequencing project of 83 tetraploid potato varieties. Because of the high SNP density in potato only a limited number of SNPs is suitable for assay development on a SNP array. An obvious outcome is that flanking SNPs contribute to assay failure, particularly for assays with SNPs located in introns. We used fitTetra software to cluster the distribution of captured signals of each marker into the expected five genotypic classes (nulliplex, simplex, duplex, triplex, quadruplex), resulting in a dataset with 14,530 SNP markers. Subsequently the genotypic data obtained with the SolSTW array was used to characterize a set of 569 potato varieties, advanced breeding clones and progenitors. This resulted in the identification of several footprints of potato breeding. Firstly SNPs were dated i.e. the year of market release of the first variety showing polymorphism for a SNP locus is an indication of the ancestry of a SNP. In such a way we identified SNPs with an ancestry tracing back to heirloom varieties, and SNPs (post-1945 SNPs) tracing back to wild species used in modern introgression breeding. Secondly, the changes in allele frequency were calculated over time. Most SNPs show a relative stable allele frequency over time, and very limited genetic variation is removed from the gene-pool of potato i.e genetic erosion is almost absent. Therefore we conclude that 100 years of breeding has not been able to get rid of non-beneficial genetic variation. Only a limited number of SNPs show a rapid increased in allele frequency, which can be explained by positive selection for disease resistance by breeders, or the more frequent use of several founders.

Better understanding of the genome wide decay of Linkage Disequilibrium (LD) and population structure offers relevant knowledge to perform and interpret the results of a genome wide association study (GWAS) (**Chapter 3**). Linkage disequilibrium (LD) is a complex phenomenon, and the influence of the factors shaping LD in tetraploids is hardly studied. Therefore we used simulated data to disentangle and therewith understand

often-confounded factors underlying LD-decay. We simulated datasets differing in number of haplotypes in a population, and differing in percentage of haplotype specific SNPs. In these simulations we observed that the choice of an estimator of LD-decay has a major effect on the outcome of an LD-decay estimate, while the true LD-decay remains the same. Based on the simulation we conclude that a 90% percentile and a so-called $D_{1/2}$ (the distance where 50% of the initial LD is decayed) performed best to estimate and compare LD-decay in potato. To understand the various aspects of LD-decay in the variety panel of 537 varieties, the panel was subdivided in several groups based on the age of a variety and the population structure groups. This resulted in the identification of LD-decay over time, i.e in relatively young varieties the average size of the LD-blocks is smaller. The differences between subpopulations were smaller and are most likely the effect of the population structure. We also observed that there are very long LD-blocks caused by introgression breeding and that different a priori MAF-thresholds also can influence the outcome of LD-decay estimation.

Having both LD-decay and population structure defined a genome wide association study (GWAS) was conducted (**Chapter 4**). For this purpose αsolanine and αchaconine were measured in potato tubers. Subsequently the sum of both (total SGA) and the ratio between the two were used to discover QTLs for these traits in a GWAS. Additionally we used three bi-parental populations to validate the GWAS results. Total SGA content was confounded with population structure and therefore it was difficult to explain all phenotypic variation with SNP markers. Two QTLs (*Sgt1.1* and *Sgt11.1*) were identified which could be validated in one of the segregating populations. The ratio between α-solanine and α-chaconine was not confounded with population structure, resulted in the identification of two major-effect QTLs (*Sgr7.1 & Sgr8.1*) located near the candidate genes *SGT1* and *SGT2*, which are known for being responsible in the final steps towards either α-solanine or α-chaconine. The QTL *Sgr8.1* could be validated, however similar phenotypes were explained by different haplotypes in two populations. We show that population structure, low frequent alleles and genetic heterogeneity may explain to some degree the missing heritability in GWAS in potato.

In **Chapter 5** we describe how the method of graphical genotyping, which is widely used in diploid bi-parental populations, can be applied in a variety panel of tetraploid varieties. We show that a few discrete filtering steps in Excel can be used to display patterns that are visual representations of introgression segments and the locations of historical recombination events. Using this method we identified introgression segments from *Solanum vernei* including the *Gpa5* locus on chromosome *5* and *Solanum stoloniferum* introgression segment including a gene involved in resistance to *Potato Virus Y* on chromosome *11*. This method requires that the haplotypes that cause the phenotypic effect have to be identical by descent (IBD).

In the final chapter 6 the results of chapter 2 to 5 are discussed. We look forward on how our results can be used in future research and applied in marker-assisted breeding. Additionally some new GWAS results are presented for tuber flesh colour, foliage maturity and resistance to *Globodera pallida* pathotype 3.

Dankwoord

Na 4 jaar Wageningen en 1,5 HZPC is ie dan eindelijk af. De laatste lootjes waren zwaar, maar ik heb het gehaald. De 4 jaar in Wageningen waren een mooie periode. Een zeer prettige werkomgeving, ik heb er erg veel geleerd en deze jaren hebben mij gemaakt tot de wetenschapper/veredelaar die ik nu ben. Aan deze periode hebben een hoop mensen hun steentje bij gedragen, echter een aantal in het bijzonder.

Ten eerste, Herman, het moment dat jij me belde met de vraag of ik de komende twee jaar wat te doen had was het begin. Een theoretisch onderzoek had wat praktische handen nodig, daarvoor vond je mij een geschikte kandidaat. Echter al vrij snel kwam de 20K array, waarover dit proefschrift gaat, ten tonele. Deze array bracht zoveel nieuwe en leuke informatie aan het licht dat we begonnen te denken aan een promotieonderzoek. Bij het traject dit van de grond te krijgen heb jij een belangrijke rol gespeeld. Samenvattend denk ik dat ik zonder jou nooit had kunnen promoveren. Ook heb je een belangrijke rol gespeeld in mijn ontwikkeling als wetenschapper, de discussies die we geregeld voerder vond ik altijd erg prettig, vaak gingen ze over wetenschap, maar ook gingen de gesprekken over onze studententijden die wat gelijkenissen vertoonden. Ik heb onze samenwerking altijd erg prettig gevonden, vooral omdat we beiden konden zeggen wat we vonden.

Zo zijn er meer collega's die een steentje hebben bij gedragen. Joao, je hebt een belangrijk rol gespeeld in het omzetten van (de vele) ideeën naar scripts in GenStat. Zonder deze vertaling was het nooit gelukt de analyses te doen die ik gedaan heb. Heel erg bedankt.

Thijs, de afgelopen jaren hebben we veel samen opgetrokken, we zagen elkaar vrij geregeld bij de koffiemachine, voor inhoudelijke discussies en andere "diepgaande" gesprekken. De laatste periode in 2014, toen ik al in Groningen woonde, heb ik bij jou en Emilie mogen logeren, erg veel dank daarvoor.

Ook Maarten heeft een belangrijk bijdrage geleverd aan een zeer gezellige periode op plantenveredeling. Menig tripje richting Veenendaal met Thijs en Andres waren een groot succes.

Ronald, het was altijd prettig om naast de wetenschappelijke zaken van gedachten te wisselen over de praktische aardappelveredeling.

Richard en Fred vooral dank voor de laatste periode waarin jullie beiden zeer efficiënt en constructief hebben bijgedragen aan de afronding van dit proefschrift.

Ook de bedrijven Agrico, Averis, HZPC, KWS en Meijer die na 2 jaar CBSG het vertrouwen in mij hebben getoond door het project met 2 jaar te verlengen. Mariëlle, Sjefke, Nick, Jan, Jacqueline, Jeroen, Emmet, Guus en Jan-David, dank voor de input tijdens onze vergaderingen, ik heb de project bijeenkomsten altijd als erg prettig ervaren. Ik hoop dat de resultaten die niet in het proefschrift staan bij kunnen dragen aan de ontwikkeling van nieuwe rassen.

Dan zijn er natuurlijk een aantal mensen op het thuisfront. Pap, bedankt voor overbrengen van de passie voor aardappels, zonder onze gedeelde passie had ik hier niet gestaan. Annemiek, sorry voor alle tijd die ik het laatste half jaar niet met jou en de jongens heb kunnen doorbrengen om proefschrift af te kunnen maken. Dank voor je begrip en ik zal deze “verloren” tijd weer inhalen.

Over de auteur

Peter Vos was born on March 17th 1983 in Dronten. In 2001 he completed high school, the Almere College in Kampen. That same year he moved to Wageningen and started his study plant sciences at Wageningen University. In June 2010 he obtained his master degree in plant sciences in the specialization Plant Breeding and Genetic resources. In November 2010 he started to work for the laboratory of plant breeding as researcher on potato within the framework of the Centre of Biosystems and Genomics (CBSG). After 2 years this was converted into a PhD research project under the supervision of Dr. ir. Herman J. van Eck, Prof Dr. Fred A. van Eeuwijk and Prof. Dr. Richard G. F. Visser. The research focused on genome wide association mapping in tetraploid potato using a 20K SNP array. The results of this research are presented in this PhD thesis. From January 1st 2015 he is working as potato breeder at HZPC Holland BV, Metslawier, The Netherlands.

Education certificate

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: Peter G. Vos
Date: 17 November 2016
Group: Laboratory of Plant Breeding
University: Wageningen University & Research

1) Start-up phase	date
► **First presentation of your project**	
Title: Genome wide association mapping in potato: Towards validation of QTLs	Jan 21, 2013
► **Writing or rewriting a project proposal**	
Title: Genome Wide Association Study (GWAS) to identify QTL involved in tuber quality traits of tetraploid potato	Dec 2012
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*2.0 credits**

2) Scientific Exposure	date
► **EPS PhD student days**	
EPS PhD Student Day, Leiden, NL	Nov 26, 2013
EPS PhD Student Day, Soest, NL	Jan 29-30, 2015
► **EPS theme symposia**	
EPS Theme 4 Symposium 'Genome Biology', Wageningen University	Dec 10, 2010
EPS Theme 4 Symposium 'Genome Biology', Wageningen University	Dec 09, 2011
EPS Theme 4 Symposium 'Genome Biology', Radboud University Nijmegen	Dec 07, 2012
EPS Theme 4 Symposium 'Genome Biology', Wageningen University	Dec 13, 2013
► **Lunteren days and other National Platforms**	
Annual Meeting 'Experimental Plant Sciences', Lunteren (NL)	Apr 04-05, 2011
Annual Meeting 'Experimental Plant Sciences', Lunteren (NL)	Apr 02-03, 2012
Annual Meeting 'Experimental Plant Sciences', Lunteren (NL)	Apr 22-23, 2013
► **Seminars (series), workshops and symposia**	
Invited seminars Plant Breeding	2010-2014
CBSG meeting 2011	Feb 01, 2011
CBSG meeting 2012	Mar 01, 2012
CBSG meeting 2013	Feb 12, 2013
PBR Research Day	Feb 28, 2012
Symposium 'Omics advances for Academia en Industry: Towards true molecular plant breeding'	Dec 11, 2014
► **Seminar plus**	
► **International symposia and congresses**	
EAPR meeting, Wageningen (NL)	Jun 27-30, 2010
EAPR meeting, Oulo (Finland)	Jul 24-29, 2011
Wageningen 100 Years Congress 'Next Generation Plant Breeding', Ede (NL)	Nov 11-14, 2012
EAPR meeting, Brussels, Belgium	Jul 06-11, 2014
► **Presentations**	
Poster: CBSG 2011	Feb 01, 2011
Talk: PBR Research Day	Feb 28, 2012
Poster: CBSG 2012	Mar 01, 2012
Talk: CBSG 2012	Mar 01, 2012
Talk: ASC (aardappel studie club)	May 16, 2012
Talk: EPS theme 4	Dec 07, 2012
Talk: 100 years plant breeding	Nov 11-14, 2012
Poster: CBSG 2013	Feb 12, 2013
Talk: CBSG 2013	Feb 12, 2013
Talk: Applied Statistics	Oct 18, 2013
Talk: Post Graduate School	Mar 21 2014
Talk: EAPR Brussels	Jul 06-11, 2014
► **IAB interview**	
Meeting with a member of the International Advisory Board of EPS	Nov 27, 2014
► **Excursions**	
Excursion Rijk zwaan	
Subtotal Scientific Exposure	*23.9 credits**

3) In-Depth Studies	date
► **EPS courses or other PhD courses**	
Postgraduate course 'Identity by Descent', Wageningen, NL	Jul 03-06, 2012
Summer School Natural Variation, Wageningen, NL	Aug 21-24, 2012
Summer School Bioinformatics, Wageningen, NL	Aug 26-30, 2013
► **Journal club**	
► **Individual research training**	
Subtotal In-Depth Studies	*3.9 credits**

4) Personal development	date
► **Skill training courses**	
Scientific Writing	Dec 2013
Competence Assessment	Dec 2013
Meer Academici voor de klas (Acklas)	Nov 2012-Jan 2013
► **Organisation of PhD students day, course or conference**	
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*3.6 credits**

TOTAL NUMBER OF CREDIT POINTS*	**33,4**

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.*

This research was funded by the Centre of Biosystems and Genomics and the breeding companies Agrico, Averis, HZPC, KWS and Meijer.

Cover Design : Thomas van den Berg