Identifying novel diagnostic SNP markers for potato (*Solanum tuberosum* L.) tuber starch and yield by association mapping

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Elske Maria Schönhals

aus Kirchheimbolanden

Köln, 2014

Die vorliegende Arbeit wurde am Max-Planck-Institut für Pflanzenzüchtungsforschung in der Abteilung für Pflanzenzüchtung und Genetik (Direktor Prof. Dr. Maarten Koornneef) angefertigt.

Berichterstatter	PD Dr. Christane Gebhardt
(Gutachter)	Prof. Dr. Martin Hülskamp

Tag der mündlichen Prüfung 28.05.2014

Science cannot solve the ultimate mystery of nature.
And that is because, in the last analysis, we ourselves are part of nature and therefore part of the mystery that we are trying to solve.

Max Planck (1932)

Abstract

The starch of potato (*Solanum tuberosum* L.) tubers is a renewable resource and an important component of multiple food and non-food products. Optimized starch yield, the product of tuber starch content and tuber yield, is therefore the central selection criterion in breeding programs for starch potatoes. The aim of this work was the detection of diagnostic single nucleotide polymorphism (SNP) markers for starch yield optimization by marker-assisted selection.

A novel association mapping population of 282 potato genotypes formed the basis of this thesis. Within a collaborative project with breeders, this population was assembled and phenotyped in replicated field trials in Northern Spain for the starch yield determining traits tuber starch content, tuber yield, weight and number. The population was genotyped for known diagnostic PCR markers, SSR markers and novel SNPs in candidate genes, which were reported in the literature to have a function in starch or yield accumulation in potato or other organisms. In addition, three subpopulations were selected based on the highest (cases) and lowest (controls) trait values for tuber starch content, tuber yield and starch yield. Ninety varieties and breeding clones in total were genome-wide genotyped for 8,303 SNPs using the SolCAP Potato Array. SNPs with highly significantly different allele frequency between the case-control subpopulations for each trait were selected and genotyped in the entire population. Furthermore, the same case-control subpopulations were genotyped with next-generation RAD sequencing. Highly significant SNPs differing between cases and controls were analyzed for effects on the protein sequence and location in previously known candidate genes.

Using a mixed linear model including population structure and kinship for association analysis, 21 diagnostic SNP markers and one insertion-deletion polymorphism were identified in candidate genes for tuber starch and yield-related traits. These associations resulted from the targeted candidate gene approach as well as from the genome wide case-control study by SolCAP Potato Array genotyping. A set of 430 novel and non-obvious candidate loci for tuber starch content, yield and starch yield was obtained from the RAD sequencing approach. Nine loci were detected in both genome-wide genotyping methods and are of special interest for further analysis.

All three applied concepts resulted in the detection of novel marker-trait associations and candidate genes. This study shows the value of combining a knowledge-based associa-

tion mapping approach with genome-wide discovery of polymorphisms as a tool for the detection of novel and non-obvious candidate genes and markers.

Zusammenfassung

Kartoffelstärke ist ein nachwachsender Rohstoff und ein wichtiger Zusatzstoff in der Lebensmittelindustrie sowie ein Additiv in anderen Industriezweigen. Optimierter Stärkeertrag, das Produkt aus dem Stärkegehalt und dem Gesamtertrag der Knollen, ist daher das zentrale Selektionsmerkmal in Züchtungsprogrammen für Stärkekartoffeln. Das Ziel dieser Arbeit war die Identifizierung von diagnostischen Single Nucleotide Polymorphism (SNP) Markern um die Marker-gestütze Selektion auf optimalen Stärkeertrag zu ermöglichen.
Grundlage dieser Arbeit war eine neue Assoziationskartierungspopulation von 282 Kartoffel-Genotypen. Diese wurde im Rahmen einer Zusammenarbeit mit Züchtern zusammengestellt und für Merkmale phänotypisiert, die den Stärkeertrag bestimmen, nämlich Stärkegehalt und Gesamtertrag der Knollen, sowie Knollen-Anzahl und Gewicht. Die Population wurde mit diagnostischen PCR Markern aus früheren Studien, SSR Markern und neuen SNPs aus Kandidatengenen, die in der Literatur beschrieben sind, genotypisiert. Zusätzlich wurden aus der Gesamtpopulation drei Subpopulationen zusammengestellt, die aus je zwei Gruppen von Genotypen mit den höchsten (Fälle) bzw. niedrigsten (Kontrollen) Werten für Knollenstärkegehalt, Ertrag und Stärkeertrag bestanden. Die insgesamt neunzig Sorten und Zuchtklone wurden mittels des genomweiten SolCAP Kartoffel SNP Array (8.303 SNPs) genotypisiert. SNPs mit hochsignifikant unterschiedlicher Allelhäufigkeit zwischen den Fall- und Kontrollgruppen für das jeweilige Merkmal wurden in der Gesamtpopulation genotypisiert. Die gleichen Subpopulationen wurden darüber hinaus mittels Next-Generation RAD-Sequenzierung genotypisiert. Signifikante SNPs zwischen der Fall- und Kontrollgruppe wurden auf ihre Auswirkung auf die Proteinsequenz durch nicht synonyme SNPs sowie auf ihre Präsenz in bisher bekannten Kandidatengenen analysiert.
Mit Hilfe eines linearen gemischten Modells, das Populationsstruktur und Verwandschaftsgrad berücksichtigte, wurden 21 diagnostische SNP Marker und ein Insertions-Deletions Polymorphismus für Knollenstärke- und Ertragsmerkmale identifiziert. Diese Assoziationen resultierten aus dem Kandidatengen-Ansatz sowie der genomweiten Genotypisierung anhand des SolCAP Kartoffel SNP Arrays. Mittels RAD Sequenzierung wurden 430 neue und nicht-offensichtliche Kandidatengene für Knollenstärkegehalt, Gesamtertrag und Stärkeertrag gefunden. Neun dieser Kandidatengene resultierten aus beiden genomweiten Genotypisierungsmethoden und sind von besonderem Interesse für weitere Analysen.
Alle drei Konzepte dieser Arbeit führten zur Identifikation neuer Marker-Merkmal Assozi-

ationen und Kandidatengene. Diese Studie zeigt, dass die Kombination eines wisssensbasierten Assoziationskartierungs-Ansatzes und genomweiter SNP Analyse in Fall-Kontroll-Studien einen wichtigen Beitrag zur Erkennung neuer, nicht-offensichtlicher Kandidatengene und Marker führt.

Contents

1. General Introduction

1.1. The potato success story

Potatoes have been an important food crop for a long time. They are a valuable source of carbohydrates, proteins and vitamins as well as minerals. The beginning of the cultivated potato was most likely a single domestication event originating from *Solanum bukasovii* in southern Peru (Spooner et al., 2005). The potato plant has taken roots in Europe between 1570-1572 when missionaries brought it to Spain upon their return back from South America. In the beginning, the plant was acknowledged for its pretty flowers. The first documented mention of the potato as food was in a letter in 1577 (Oliemans, 1988). The nun Teresa of Ávila expressed her appreciation for a parcel of potatoes and their good taste. Within three centuries after the introduction of potato to Spain, the plant made its way through Europe and struck roots deeply into European culture. About 300 years later, Vincent van Gogh depicted a typical peasants meal in his famous painting *The Potato Eaters* (1885): five people are sitting around a table, eating potatoes from one large dish (Rijksmuseum Kröller-Müller, 1961).

Potato applications are manifold In former times, the potato was "the bread of the poor" (Oliemans, 1988), but nowadays it is an important player in the diets of cultures all over the world. It ranges amongst the top five agricultural crops - after corn, wheat and rice - with an average world potato production of 370 million tons per year (FAO Crops Statistics Database, 2013).
Despite the importance as a food crop, not all grown potatoes are designated for consumption. About 10 million tons of potatoes were harvested in Germany in 2010. Most of these potatoes (70%) were consumed as table potatoes or processed products like chips and french fries. The second largest share of 20% was used for the production of starch by the starch industry (Statistisches Bundesamt, 2012). Potato starch is extracted from tubers and applied as an additive in several industrial processes. Depending on the application, starches for industrial use are often chemically modified (Ellis et al., 1998). The total starch production in Europe in 2010 was 10 million tons. 14% of this starch was produced from potatoes, whereas in Germany this proportion amounted to nearly 50% of the total starch production in 2010 (Fachverband der Stärke-Industrie, 2013). Consequently, the

potato plays an important role in the starch industry.

Potato starch Potato starch has some features that makes it different from corn or wheat starch. It has a neutral taste, a white color, a higher viscosity and a higher purity with low lipid and protein content (Jobling, 2004; Südstärke Informationsdienst, 2013a). The major industrial sectors for potato starch are the food, paper, general and textile industries (Avebe, 2013; Ellis et al., 1998). In the food industry, potato starch is applied as an additive for example in flour instant mixes, custards, packet soups or infant formulae (Ellis et al., 1998; Südstärke Informationsdienst, 2013c). Major applications in the paper industry are in paper production, surface coating and the rubber lining of, for example, envelopes (Südstärke Informationsdienst, 2013d). Furthermore, it is applied in the production of textiles in weaving mills (Südstärke Informationsdienst, 2013e) and for a multitude of applications in general industry, like fertilizer granulation and as binder for fiberglass fabrics (Avebe, 2013; Südstärke Informationsdienst, 2013b).

1.2. Potato genetics

The cultivated European potatoes *Solanum tuberosum* (ssp. *tuberosum*) are autotetraploids (2n=4x=48), which means that they have four alleles per locus. Homologous chromosomes pair at random during meiosis (Milbourne et al., 2007). In addition, there are tuber-bearing varieties under cultivation that are non-*tuberosum* species with ploidy levels ranging from diploid to hexaploid (Van den Berg and Jacobs, 2007). Potatoes are outbreeding plants. Therefore they obtain a high level of heterozygosity and are prone to inbreeding depression, making it difficult to obtain homozygous lines. The heterozygosity in commercial cultivars is preserved by the clonal propagation of tubers (Milbourne et al., 2007; The Potato Genome Sequencing Consortium, 2011).

1.3. Marker technology

The molecular marker technologies all derive from the natural DNA sequence variation of individuals (Gebhardt, 2005). A multitude of markers has been developed for the visualization of this variation over the last decades. The first molecular markers were developed based on polymorphisms in restriction enzyme recognition sites (RFLP; Botstein et al., 1980). They were followed by PCR-based marker systems, for example microsatellite (SSR) markers, RAPD markers (Welsh and McClelland, 1990; Williams et al., 1990), AFLP markers (Vos et al., 1995), single-strand conformation polymorphism (SSCP) markers or cleaved amplified polymorphic sequences (CAPS) markers.

Sequencing technologies, such as Sanger amplicon sequencing, then allowed for the detection of single nucleotide polymorphisms (SNPs) and therefore the estimation of allelic dosages at SNP loci for individual genotypes. In potato, five allele combinations can be present at a bi-allelic SNP position: two homozygous combinations (nulliplex, quadruplex) and three heterozygous combinations (simplex, duplex and triplex).

SolCAP Potato Array Custom-made arrays are used for the analysis of large amounts of SNPs per genotype (e.g. Hamilton et al., 2011). A recent development for potato in this field is the Infinium 8303 Potato Array (Felcher et al., 2012), hereafter referred to as the SolCAP Potato Array. The SolCAP Potato Array allows the simultaneous genotyping of an individual with 8,303 SNP markers. It was developed from sequence information of five North American (Atlantic, Premier Russet, Snowden, Kennebec, Shepody) and one European potato variety (Bintje). Sequence data was generated by next-generation transcriptome sequencing in the case of Atlantic, Premier Russet and Snowden as well as drawn from public databases for Bintje, Kennebec, Shepody (Hamilton et al., 2011). The marker loci for the array were selected in a way, that 36% of the markers are located in candidate genes of interest and 6% are previously mapped markers. Further markers (57%) in coding regions were selected for a maximum coverage of the potato genome (Felcher et al., 2012). So far, three papers have been published implementing the SolCAP Potato SNP Array. Felcher et al. (2012) describe the design of the array as well as the integration of two diploid linkage maps with the potato genome sequence. Also a tetraploid bi-parental mapping population was genotyped with the SolCAP Potato Array (Hackett et al., 2013). Its first application in European potato germplasm was in a pilot study by Stich et al. (2013), where population structure and linkage disequilibrium were tested in a set of 36 tetraploid cultivars and eight diploid potato clones.

RAD sequencing The progress of next-generation sequencing (NGS) methods and the decreasing prices for sequence runs have led to a number of novel techniques for the detection of polymorphisms between genotypes. Many methods reduce the complexity of the genome. Some examples are genotyping by sequencing (GBS; Uitdewilligen et al., 2013), sequence-based genotyping (SBG; Truong et al., 2012) and restriction-site associated DNA tag sequencing (RAD sequencing; Baird et al., 2008). RAD sequencing reduces the size of the genome by a restriction enzyme fragmentation and a subsequent shearing and size selection step. In combination with NGS it generates sequences of RAD tags that are distributed over the whole genome. The method was first described by Baird et al. (2008) and a detailed protocol was provided by Etter et al. (2011). The huge amount of genetic markers that is gained in comparison to the relatively low cost for library preparation and NGS makes the method attractive also for non-model species. As a result of this, the examples

for the implementation of RAD sequencing are already manifold and the list still keeps expanding. The following is only a representation of the most prominent examples. The applications reached from fine-mapping and population genetics in threespine stickleback (Baird et al., 2008) and guppy (Willing et al., 2011) to marker detection for germplasm-genotyping in bornean elephant (Sharma et al., 2012), globe artichoke (Scaglione et al., 2012) and eggplant (Barchi et al., 2011). So far eggplant is the only representative of the solanaceous crops that has been genotyped by RAD sequencing. Furthermore, genotyping mapping populations by RAD sequencing for constructing high-density linkage maps was used in barley (Chutimanitsakun et al., 2011), perennial ryegrass (Pfender et al., 2011) and grapevine (Wang et al., 2012). Until now, the utilization of RAD sequencing has also been described for genotyping of a number of tetraploid species, like rapeseed (Bus et al., 2012), bamboo (Wang et al., 2013) and sturgeons (Ogden et al., 2013). The special strength of RAD sequencing is its high-throughput polymorphism detection.

1.4. Linkage analysis and physical map of potato

1.4.1. Genetic maps of potato

The construction of genetic maps of potato is based on the principle of linkage analysis. Two loci that lie in physical distance to each other on the same chromosome are considered to be genetically linked as they are not inherited independently (Griffiths et al., 2005). For the construction of a linkage map, a mapping population is essential. The offspring should be segregating for polymorphic markers. By calculating the recombination frequency, the genetic distance between loci can be estimated (Gebhardt, 2007).

The first genetic linkage maps of potato were based on RFLP marker technology and used the synteny between potato and tomato. Bonierbale et al. (1988) presented the first genetic map of potato based on an interspecific cross between diploid potato lines. The first map within the gene pool of diploid *S. tuberosum* was released by Gebhardt et al. (1989). These genetic maps function as reference molecular maps. Many more linkage maps were published in the years following the first publication of genetic maps of potato (e.g. Gebhardt et al., 1991; Milbourne et al., 1998; Van Eck, 1995).

To close the gaps between markers in the same linkage groups, an ultra-high density genetic map was established based on 10,000 polymorphic AFLP markers, which is considered the most elaborate linkage map of potato (Van Os et al., 2006).

In a recent study, Hackett et al. (2013) genotyped an intensively studied tetraploid mapping population (Bradshaw et al., 2008; Meyer et al., 1998) with the SolCAP Potato Array (Felcher et al., 2012) and created a high-density linkage map based on approximately 4000 polymorphic SNPs.

1.4.2. Physical map of the potato genome

The first physical map of potato was published in 2011 (The Potato Genome Sequencing Consortium, 2011). A doubled monoploid, *Solanum tuberosum* group Phureja DM1-3 516 R44, hereafter referred to as the potato genome sequence, was sequenced. Since the initial release, the sequence has been updated continuously. Version v4.03 is the most recently improved version that was constructed by integrating information from genetic and physical maps (Sharma et al., 2013). With the availability of the potato genome sequence it is now feasible to compare linkage maps with the physical positions of markers and to look at QTL regions in more detail. Also *in silico* mapping of candidate genes by sequence homology and the estimation of the copy numbers are now possible.

1.5. QTL linkage and association mapping in potato

The major perspective of genetics is the correlation of phenotypic variation with DNA sequence variation (Van Eck, 2007). Two leading concepts are available to achieve this: quantitative trait locus (QTL) linkage mapping, which is based on genetic maps, and association mapping, which is based on linkage disequilibrium (LD).

1.5.1. Linkage mapping

The association of phenotypic trait values with segregating alleles of molecular markers in a mapping population is referred to as linkage mapping. The aim of linkage mapping is to detect genomic regions that explain phenotypic variation in a trait of interest and the subsequent identification of potential causal genes in that region. QTL are regions on the chromosomes which are physically linked to a molecular marker allele. The QTL and the marker allele are inherited together. Underlying genes of a quantitative trait, which has a wide distribution of phenotypes, can be located on all chromosomes (Gebhardt et al., 2005). For linkage analysis, several types of mapping populations are suitable (Collard et al., 2005). After establishing the mapping population, it is genotyped with segregating molecular markers and phenotyped for the quantitative trait of interest. A linkage map is produced from the molecular marker data and QTL are detected by marker-trait association.

QTL linkage analysis in potato is mainly carried out on diploid level. This is due to the highly heterozygous nature of the potato plants. A large number of QTL studies deal with resistances to biotic stresses like *Phytophthora infestans* (e.g. Li et al., 1998), root cyst nematodes (e.g. Kreike et al., 1994) and abiotic stresses (e.g. drought tolerance: Anithakumari et al., 2011). Furthermore, yield- and quality-related traits were studied with QTL mapping, such as specific gravity (Freyre and Douches, 1994), starch content and

yield (Schäfer-Pregl et al., 1998), cold-sweetening (Menéndez et al., 2002) and enzymatic discoloration (Werij et al., 2007).
Although linkage mapping in tetraploid potato is not as straight-forward as in diploid potato, there are successful examples, such as the resistance studies for late blight (Bradshaw et al., 1998; Li et al., 1998; Meyer et al., 1998). Bradshaw et al. (2008) mapped 16 QTL for yield, agronomic and tuber quality traits in a tetraploid full-sub family mapping population. More examples were reviewed by Milbourne et al. (2007) and Van Eck (2007).

1.5.2. Association mapping

Alternatively to the family-based linkage mapping approach, association mapping is a method for the detection of marker-trait associations in a population of individuals that are related by descent. The method takes advantage of historical meiotic recombinations and linkage disequilibrium (Flint-Garcia et al., 2003). It was first established in the study of complex inherited diseases in human populations, where it is not feasible to establish segregating mapping populations from crosses (Gebhardt et al., 2004).
For association mapping, a population consisting of diverse germplasm including cultivars, breeding clones and landraces is assembled and phenotyped for the complex traits of interest. Molecular markers are then analyzed in the population and marker-trait associations between phenotypic and genetic variation are detected. In the case of candidate gene association mapping, the molecular markers are obtained from knowledge-based candidates, whereas markers for genome-wide association mapping randomly cover all chromosomes in high density.
Association mapping is based on linkage disequilibrium (LD). LD is defined as the non-random association of two alleles in a population (Flint-Garcia et al., 2003). This is often the case for loci that are in close proximity to each other on the same chromosome (linkage). However, LD can also occur between alleles on different chromosomes (Flint-Garcia et al., 2003). There are different opinions regarding the extend of LD in tetraploid potato. D'hoop et al. (2010) report a distance of 5 cM for genome-wide LD. Stich et al. (2013) suggest a linkage decay within 275 bp. Association mapping is an application of LD (Soto-Cerda and Cloutier, 2012), where the associated marker and the quantitative trait locus are in LD or physically linked in the ideal case (Gebhardt, 2013).

Population structure analysis to avoid false-positives in association mapping The genotypes of a potato population are a collection of individuals that are related by descent (Gebhardt et al., 2005). Consequently, there is a potential bias towards relatedness in the statistical analysis, which means that a trait of interest can, for example, be linked to a gene pool or a geographic origin (Flint-Garcia et al., 2003). The information about the

degree of relatedness between genotypes in the mapping population plays a critical role in association mapping in order to avoid false positives. While a marker may not be linked to a QTL, there is a considerable risk of detecting a significant association only based on the genetic relatedness between individuals (Pritchard et al., 2000).
There are several options to assess population structure in potato based on genetic markers. The two options arising from a factor analysis approach are principal coordinate (D'hoop et al., 2010; Pajerowska-Mukhtar et al., 2009; Urbany et al., 2011) and principal component analysis (D'hoop et al., 2010), where genotyping information from molecular marker data is processed. In another approach, the marker data are analyzed by Bayesian clustering, implemented in the software STRUCTURE (Pritchard et al., 2000). This has been applied in the field of potato research in several studies (D'hoop et al., 2010; Li et al., 2008; Pajerowska-Mukhtar et al., 2009; Simko, 2004; Simko et al., 2006). Further options for population structure assessment are Analysis of Molecular Variance (AMOVA) and hierarchical clustering (D'hoop et al., 2010).

Kinship In an association mapping population, substructure can be present caused by identity by descent. To detect allele identity in state, a kind of relatedness independent from identity by descent (Stich et al., 2008), a kinship matrix (Yu et al., 2006), is required for the analysis. The kinship matrix includes the pairwise comparison between all genotypes based on molecular markers. Stich et al. (2008) prefer this method to the use of pedigree-based kinship (Malosetti et al., 2007). Several methods for estimating kinship are available, for example, the method of Loiselle et al. (1995), VanRaden (2008) or the EMMA algorithm (Kang et al., 2008).

Applications of association mapping Association mapping has been used to detect genetic variation that explains variation of complex traits in plants, as for example in corn (Wilson et al., 2004), wheat (Breseghello and Sorrells, 2006), barley (Cockram et al., 2008) rice (Huang et al., 2011), perennial ryegrass (Skot et al., 2005), *Arabidopsis* (Aranzana et al., 2005), rapeseed and sugar beet (Stich and Melchinger, 2009). The first example of association mapping in tetraploid potato germplasm was published by Gebhardt et al. (2004), who studied an assembled collection of 600 potato cultivars to detect marker-trait associations for late blight resistance and maturity based on historic recombination events. Further association mapping studies based on candidate genes followed for resistance against *Verticillium dahliae* (Simko et al., 2004) and *Phytophthora infestans* (Malosetti et al., 2007; Pajerowska-Mukhtar et al., 2009). More examples are yield and tuber quality traits such as tuber starch content, tuber yield, starch yield and chip quality (Fischer et al., 2013; Li et al., 2008, 2005). Similarly, tuber bruising susceptibility, tuber shape and plant maturity were studied by association mapping in tetraploid potato

(Urbany et al., 2011).
A broader way of looking at marker-trait associations is by genome-wide association mapping. D'hoop et al. (2008) gave a first example of this approach, although the amount of markers used in the study was still rather low. Another example for genome-wide association mapping in a small genotype panel was described by Uitdewilligen et al. (2013). There are no further examples of genome-wide association mapping in potato.

1.5.3. Advantages of association mapping compared to linkage mapping

According to Flint-Garcia et al. (2003), there are three main advantages of association mapping compared to linkage mapping.
Firstly, the mapping resolution of association mapping is better due to the higher amount of meiotic events, whereas linkage mapping generally looks at the recombination in a single meiotic generation (Gebhardt, 2007). However, when working with potatoes, this is not such a significant advantage, since Gebhardt et al. (2004) found that only relatively few meiotic generations separate individual genotypes. This is likely due to the clonal propagation of potato whereby the meiotic generation is conserved.
Secondly, a high number of alleles can be detected with association mapping. In a segregation population, the maximum amount of different alleles possibly detected at one locus in the offspring of a diploid linkage mapping population are four and eight in a tetraploid linkage mapping population. In an assembled population of 200 tetraploid genotypes, the theoretical maximum number of different alleles at one locus is 800. Because of a reduced statistical power, marker-trait associations of very rare alleles are not likely to be detected. Therefore, association mapping is mainly suitable for the detection of common variants (Flint-Garcia et al., 2003).
Thirdly, the markers can be immediately applied in breeding programs. Detected markers are directly and broadly applicable when the mapping population consists of appropriate breeding material (Li et al., 2013; Stich and Melchinger, 2010).

1.6. Pooling strategies for dissecting the genetic background of quantitative traits

Although the prices for high-throughput genotyping have decreased a lot with the development of novel high-throughput genotyping methods, such as the SolCAP Potato Array, it is still a costly procedure for breeders to analyze their complete germplasm. Pooling strategies have been developed for reducing the amount of genotypes in analyses. A prominent concept in plants is bulked segregant analysis and a concept from human disease studies is the case-control study.

1.6.1. Bulked segregant analysis

Bulked segregant analysis requires a segregating bi-parental mapping population (Collard et al., 2005). Genotypes with extreme phenotypes are bulked into two groups and are then compared at the genetic level for the identification of the underlying genetics. The focus of bulked segregant analysis is on the detection of markers closely linked to quantitative trait loci (Collard et al., 2005; Meksem et al., 1995). A bulked segregant analysis pooling strategy can be a shortcut to the identification of markers that tag quantitative trait loci (Collard et al., 2005) and it is an effective screening approach for the detection of candidate genes that have an effect on the traits of interest (Kloosterman et al., 2010).
The analysis of complex traits by bulked segregant analysis has been applied by a multitude of crop studies, for example in corn (Quarrie et al., 1999), barley (Chen et al., 2011), rice (Takagi et al., 2013; Zhang et al., 2009) and grapevine (Donald et al., 2002). Apart from these diploid crops, bulked segregant analysis has been used for fine-mapping of genes in polyploid wheat (Trick et al., 2012). In potato, bulked segregant analysis has been applied for qualitative traits (Li et al., 1998; Meksem et al., 1995), but also quantitative traits were studied in diploid and tetraploid genetic background. In their study on potato wart disease resistance, Ballvora et al. (2011) analyzed two tetraploid half-sib families. Similarly, Kloosterman et al. (2010) tested the suitability of bulked segregant analysis in a diploid mapping population. They aimed at the identification of candidate genes for tuber flesh quality and cooking type as well as free methionine content by conducting a bulked segregant analysis profiling experiment.

1.6.2. Case-control studies

In a case-control study, phenotypically different groups are compared to dissect the genetic background of complex traits. The concept is widely used in human disease risk studies (Balding, 2006) and therefore phenotypic groups are related by descent, rather than selected from a bi-parental mapping population as done in bulked segregant analysis. The objective is to screen these pools for obtaining an idea of the relevant disease and the risk factor as well as the genetic background of the observed differences (Balding, 2006).
Huang et al. (2011) suggested the use of a case-control study to maximize the probability of finding new marker-trait relations in corn as compared to a genome-wide association study. In potato, this approach has successfully been adopted for the identification of candidate genes by comparative proteomics. Urbany et al. (2012) have assembled two genetic pools of ten genotypes per group, which differ in tuber bruising susceptibility. By comparing the protein expression between both pools, novel candidate genes were detected. The obtained candidate genes were subsequently tested in an association mapping population of potato cultivars (Urbany et al., 2011). Similarly, Fischer et al. (2013) applied compar-

ative proteomics to identify novel candidate genes that are associated with cold-induced sweetening. The individuals in the pools showed an extremely low and an extremely high amount of studied trait, respectively. Candidate gene association mapping confirmed that DNA sequence variation explained part of the phenotypic variation of cold sweetening in potato tubers (Fischer et al., 2013).

1.7. QTL mapping of tuber quality, starch and yield-related traits in potato

Reference molecular function maps were established for resistance traits (Gebhardt and Valkonen, 2001) as well as carbohydrate metabolism, sugar accumulation and transport (Chen et al., 2001; Menéndez et al., 2002; Schäfer-Pregl et al., 1998; Werij et al., 2012) to locate candidate genes on the genetic map and to see to which extent they co-localize with QTL regions. In more recent studies, candidate gene based association mapping revealed marker-trait associations for resistance, yield and quality traits. In this section, the focus is on the tuber quality, starch and yield-related traits, which were mapped in potato.

1.7.1. Tuber and starch quality traits

Tuber quality According to Van Eck (2007), tuber quality traits are starch content, discoloration - such as processing quality, cold-sweetening, enzymatic discoloration and bruising - cooking type and texture, glycoalkaloid content, growing defects and tuber size uniformity.

Potato tuber starch content, which is defined as the percentage of starch in relation to the total tuber weight (Von Scheele et al., 1936), ranges between 15-20% (Schäfer-Pregl et al., 1998) and is of special importance for the nutritional value. Furthermore, the potato starch mills calculate the price for potatoes based on tuber starch content and tuber yield (Avebe, 2013; Mahl, 2013, personal communication). Schäfer-Pregl et al. (1998) and Werij et al. (2012) studied tuber starch content in diploid mapping populations, resulting in many QTL for the trait. Candidate gene based marker-trait associations were detected by Li et al. (2005), Li et al. (2008), Draffehn et al. (2010), Urbany et al. (2011) and Fischer et al. (2013).

Processing quality is of major importance for the production of potato chips and French fries. QTL studies for frying color (Bradshaw et al., 2008), chipping color (Douches and Freyre, 1994; Werij et al., 2012), cold-sweetening and sugar accumulation (Menéndez et al., 2002; Werij et al., 2012) were performed and many genomic regions were identified. Molecular markers for frying color (D'hoop et al., 2008) and chip quality (Fischer et al., 2013; Li et al., 2008, 2005) were detected by association mapping.

Furthermore, tuber quality traits were investigated by QTL studies (Bradshaw et al., 2008, cracks in tubers), bulked segregant analysis (Kloosterman et al., 2010, cooking type) and association mapping (Urbany et al., 2011, bruising susceptibility).

Starch quality traits Starch has very different properties, depending on its source and individual composition. It is often modified for the applications in the food and the pharmaceutical industry (Ellis et al., 1998). As mentioned previously, potato starch is well-known for its high quality and special features. Special starch quality traits mapped by linkage analysis in potato were, for example, amylose content (Van de Wal et al., 2001; Werij et al., 2012), starch phosphorylation, gelling temperature and average granule size (Werij et al., 2012). Candidate gene based association mapping resulted in applicable markers for starch phosphorylation in breeding programs (Uitdewilligen, 2012; Wolters et al., 2011).

1.7.2. Tuber yield

Probably one of the most important traits for potato growers is tuber yield. Higher yield levels result in cost optimization and potentially lower environmental impact. Most breeders automatically breed for this trait with the pre-selection of phenotypes that are well-performing. Potato yield is a very complex trait that is strongly influenced by environmental conditions, agricultural system and preceding crop (Becker and Leithold, 2008) on the one hand and genetic composition on the other hand. The average yield of potato lies at about 43 tons per hectare (Bundesministerium für Ernährung Landwirtschaft und Verbraucherschutz, 2013, data for Germany 2007-2012).
There is only a limited number of studies that deal with the mapping of yield traits. Several QTLs for yield were reported by Schäfer-Pregl et al. (1998) and Bradshaw et al. (2008). Marker-trait associations for tuber yield were detected in the association mapping studies of D'hoop et al. (2008), Li et al. (2008, 2005) and Urbany et al. (2011). There are no really obvious candidate genes for this trait and there is a lack of molecular markers that can be applied in breeding programs.

1.7.3. Starch yield

The most interesting feature of potato as a renewable resource for starch is the amount of starch that can be obtained per unit of arable land. Hence, in potato cultivation for the starch industry it is of major interest to produce the maximum possible amount of starch per plant. Starch mills pay potato farmers according to the amount of starch yield, which is the calculated product of tuber starch content and potato tuber yield. These two traits are negatively correlated (Li et al., 2013; Urbany et al., 2011).

So far, candidate gene based association mapping revealed only a few marker-trait associations for starch yield. The majority were detected and confirmed in the same genetic background of the CHIPS-ALL population (Draffehn et al., 2010; Fischer et al., 2013; Li et al., 2008, 2005).

1.8. Breeding for optimized starch yield traits

1.8.1. Marker-assisted selection in potato

Breeding efforts can be accelerated by marker-assisted selection (MAS), selecting for diagnostic molecular markers that are associated with traits of interest. A diagnostic molecular marker is linked to a trait of interest or is tagging a region that is associated with the trait of interest. A selection for these markers in the progeny of a cross enlarges the genetic basis for the traits in the offspring and is a critical factor of the marker-assisted breeding process (Collard et al., 2005; Li et al., 2013). Li et al. (2013) confirmed the applicability of allele-specific PCR-based markers for marker-assisted selection in breeding populations with different genetic background. Ortega and Lopez-Vizcon (2012) furthermore described the use of molecular markers for disease resistance in a commercial potato breeding program.

Genomic selection Compared to other crop breeding programs, the number of markers in potato used in breeding programs is rather low. In corn or wheat breeding programs, the concept of genomic prediction has been acquired, where a large amount of genome-wide genetic markers is used for "accelerating genetic gains" (Crossa et al., 2013). Genomic predictions of genotype performance in the field can be made based on the genetic markers (Lado et al., 2013). At present, potato breeding programs are far from that. The so-called *breeders eye* and phenotyping results are still the most important factors in the selection process, along with a few molecular markers.

Markers for starch yield optimization Three major traits are involved in breeding for starch yield optimization: tuber starch content, tuber yield and starch yield, since starch yield is the product of tuber starch content and tuber yield. When looking at the natural variation in the potato germplasm, there is evidence for a wide distribution of the traits. Especially for tuber starch content, where the wild *Solanum* species can have up to 50% dry matter, as compared to the cultivated *S. tuberosum* genotypes (Jansen et al., 2001). As a drawback, these wild species have poor agronomic performance and low yield. Tuber starch content is a trait that can be easily assessed and usually has a high heritability (Li et al., 2013; Urbany et al., 2011). Tuber yield, on the other hand, can only be measured

late in the breeding process, when sufficient tubers are generated for a yield assessment trial. As a result of this, the information on starch yield can only be obtained at the same time as yield. A large set of diagnostic markers, especially for tuber yield and starch yield, could significantly accelerate the breeding process for these traits.

1.8.2. Genetically modified potatoes

With genetic modification, novel genetic variation is introduced into existing potato cultivars. The most prominent example for potato breeding with the application of genetically modification is the cultivar Amflora (BASF Plant Science, Limburgerhof, Germany). Its starch is amylose-free which is achieved by an antisense construct targeting granule-bound starch synthase I.
An alternative for the introduction of novel variation is a mutagenesis approach. Plants with novel traits that are generated by this method are not considered as genetically modified organisms. Muth et al. (2008) introduced a method for precision breeding for novel starch properties. EMS-induced point mutations in an allele of granule-bound starch synthase I causes protein truncation. Plants homozygous for this mutation showed a phenotype with reduced amylose contents.
Furthermore, there are studies about the modification of transporters in the plants that lead to altered starch yield-related traits. Regierer et al. (2002) reported a large increase of tuber starch content and tuber yield by transgenic down-regulation of adenylate kinase activity. By transgenic over-expression of Glucose-6-phosphate translocator and adenylate translocator in amyloplasts, tuber yield and tuber starch content were strongly increased (Zhang et al., 2008). When not only sink capacity but also source capacity in leaves was strengthened, an almost doubled amount of tuber starch yield could be measured in transformed potato plants (Jonik et al., 2012).

1.9. Objectives of this thesis

The objective of this study was the detection of diagnostic SNP markers for starch yield optimization in potato to facilitate marker-assisted selection. The main focus lay on potato tuber starch and yield-related traits that were studied in a novel association mapping population of commercially relevant potato genotypes.
Three concepts were applied for the identification of sequence variation in the population: a knowledge-based candidate gene study in which SNPs were identified in genes of the starch metabolic pathway and marker-trait associations were detected by association mapping (Chapter 2). Furthermore, 90 genotypes in three case-control populations were genotyped with the SolCAP Potato Array (Chapter 3) as well as by next-generation RAD sequencing

(Chapter 4).
The underlying hypothesis was that sequence variation in genes can be detected by these concepts and that the identified sequence polymorphisms explain variation in phenotypic traits related to starch yield in potato.

2. Association mapping for starch yield optimization in a novel potato population

2.1. Background

The market for industrially use potatoes is a fast-growing sector, reaching 20% of the total production at present (Statistisches Bundesamt, 2012, data for Germany in 2010). Industrially used potatoes are mainly processed for the production of starch and starch derivatives. Potato starch yield - meaning the amount of starch extracted from the total tubers of one potato plant - is a trait of major interest in the potato breeding process. It is the product of tuber starch content and tuber yield per area unit (dt/ha). Tuber starch content can be easily measured and is a highly heritable trait (Bradshaw et al., 2008; Urbany et al., 2011), with several markers available for molecular marker-assisted breeding (see Chapter 1). Tuber yield and tuber starch yield can only be assessed late in the breeding process and only few molecular markers are available.

Candidate gene association mapping in potato In the preceding years, plant breeding research has adopted association mapping from human population genetics, which has the advantage that populations do not have to be generated from specific crosses (Flint-Garcia et al., 2003). Populations are assembled from the pool of existing genotypes.
Association mapping is based on the application of linkage disequilibrium (LD). LD is the non-random association of alleles in a population of individuals related by descent. Marker-trait associations can be established because the alleles of a molecular marker are in LD with QTL (Flint-Garcia et al., 2003).
Association mapping by a candidate gene approach makes use of knowledge-based candidates from literature or reference molecular function maps (e.g. Chen et al., 2001; Menéndez et al., 2002; Schäfer-Pregl et al., 1998; Werij et al., 2012). Molecular markers are then developed for these loci and marker-trait associations are identified by a suitable statistical model. For avoiding false positives in the detection of marker-traits associations in candidate gene association mapping, population structure and kinship are taken into account as part of a mixed linear model (Stich et al., 2008; Yu et al., 2006).

Approach In order to detect diagnostic SNP markers for tuber starch yield by candidate gene association mapping, a novel potato population was established and phenotyped for starch as well as yield related traits. Knowledge-based candidate genes were selected and marker-trait associations were analyzed from sequence variation in the selected loci by applying a mixed linear model that accounts for a hidden population structure and a marker-based kinship relationship.

2.2. Material and Methods

2.2.1. Plant material

A new population was established for this project. In total 350 genotypes (Appendix B, Table B.1) were planted and phenotyped at the breeding company Appacale (Burgos, Spain) and at the Neiker research institute (Vitoria-Gasteiz, Spain). Two different subpopulations of tetraploid breeding clones and varieties were grown in an open field trial at the two sites in two subsequent years. 50 standard commercial varieties were included in both subpopulations. Furthermore 18 landraces and 20 diploid individuals were planted in the trial (Table 2.1). 50 *Solanum* wild species, which were kept *in vitro*, were grown in tunnels at the Neiker research station in 2010.

Table 2.1.: **Planted potato genotypes at the Appacale and Neiker sites in the two trial years.** All genotypes were phenotyped with the exception of the Wild *Solanum* species

Location	2010	2010	2011	2011
	Appacale	Neiker	Appacale	Neiker
Standards varieties	50	50	50	50
Commercial cultivars	47	94	36	94
Breeding clones	60	13	45	13
Landraces	–	18	–	12
Diploid clones	20	–	8	–
Wild *Solanum* species	–	50	–	–
Total genotypes	177	225	139	169

2.2.2. Experimental design

The trials were planted in an Augmented Design (Petersen, 1985) in blocks of 25 cultivars with the varieties Desirée, Kennebec and Jaerla as testers in each block.
At the two sites, the experimental setup was slightly differing for each year. At the Neiker site, 10 tubers were planted per clone in one block. In 2010, four representative, single plants were harvested and phenotyped in each plot. Four plants of each plot were harvested and bulked for phenotypic analysis in 2011.
At the Appacale site, two tubers were planted for each genotype in 2010. These two plants were harvested and bulked for further phenotypic analysis. To reduce the influence of environmental variation, the number of plants grown and phenotyped was increased to six plants per plot in 2011.

2.2.3. Phenotypic data

The plants were phenotyped by the breeding companies Neiker and Appacale. The measured traits and their units were as follows: tuber yield (g/plant) (TY), tuber number (tubers/plant) (TN), tuber under water weight (g/plant). Tuber weight (g/tuber) (TW) was calculated from TY and TN. Amylose content in percent of the total starch (AMY) was measured, following the method of Hovenkamp-Hermelink et al. (1988). Tuber starch content (%) (TSC) and tuber starch yield (g/plant) (TSY) are derived parameters. Specific gravity (ρ) was calculated from TY and tuber under water weight (Equation 2.1) and TSC (Equation 2.2) was directly calculated from ρ according to Von Scheele et al. (1936). TSY is the product of TY and TSC (Equation 2.3), representing the average amount of total tuber starch in one plant. The primary phenotype data are given in Appendix B (Table B.2).

$$\rho = \frac{\text{TY (air) (g/plant)}}{\text{TY (air) (g/plant)} - \text{Tuber under water weight (g/plant)}} \tag{2.1}$$

$$\text{TSC (\%)} = 17.546 + (199.07\ (\rho - 1.0988)) \tag{2.2}$$

$$\text{TSY (g/plant)} = \text{TY (g/plant)} \cdot \text{TSC (\%)} \tag{2.3}$$

2.2.4. DNA extraction

Fresh potato leaves were sampled from all potato genotypes. The leaves were frozen in liquid nitrogen (-80°C), lyophilized and stored at -20°C. Genomic DNA was extracted from 20 mg lyophilized leaf material, using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. An extra washing step with 500 μl ethanol (96%) was performed, preceding the elution of the DNA from the colums. DNA concentrations and quality were determined using the Qubit dsDNA BR Assay Kit (Invitrogen, Karlsruhe, Germany) and visually examined by electrophoresis on a 1% agarose gel in comparison to λ DNA (Invitrogen, Karlsruhe, Germany).

2.2.5. Genotypic data

The tetraploid varieties (standard varieties, commercial cultivars), breeding clones and landraces, in total 282 individual, were genotyped with a diverse set of markers. These 282 genotypes are further referred to as the Quest population. Microsatellite markers, allele specific PCR markers, cytoplasm markers and candidate gene sequences were used for genotyping and subsequent analyses.

Microsatellite markers The results of Pajerowska-Mukhtar et al. (2009) suggested that a minimum of 25 microsatellite markers, also called SSR makers, is sufficient to identify population structure in tetraploid potatoes. Therefore, 29 SSR markers were in total selected from several sources and scored in the QUEST population (Appendix Table B.5). The main selection criteria were that the primers amplify a single locus in the genome, that the markers show clearly distinguishable bands on the gel and are polymorphic in the population. At least one marker per chromosome arm was selected.
Microsatellite markers were amplified in a 25 μl reaction volume, containing 50 ng genomic DNA, 8 mM Tris-HCL pH 8.3, 40 mM KCl, 6.4 mM $MgCl_2$, 0.08% Trifon X-100, 160 μM of each dNTP (Roth, Karlsruhe, Germany), 0.2 μM of each primer, 1U Ampliqon Taq Polymerase (Ampliqon, Odense M, Denmark) and deionized water (Merck KGaA, Darmstadt, Germany). Markers were amplified under the following PCR conditions: 3 min at 94°C, 2 min at annealing temperature (Appendix B, Table B.5), 90 sec at 72°C, followed by 29 cycles of 94°C (60 sec), annealing temperature (60 sec) and 72°C (45 sec), completed by a final elongation step of 5 min at 72°C (Provan et al., 1996). For markers with T_a 60-54°C, touchdown PCR was performed with the same procedure, lowering the temperature by 1°C per cycle until the final annealing temperature was reached. The success of the PCR and the intensity of the bands was assessed on an 2% agarose gel.
Appropriate Spreadex gels (Elchrom Scientific AG, Cham, Switzerland) for SSR allele separation were selected and running time was determined using the information of previous SSR genotyping as well as the ElQuant Software, provided online by the manufacturer. The gels were run in the Elchrom SEA 2000 system (Elchrom Scientific AG, Cham, Switzerland) using the operational instructions by the supplier. Allele sizes were measured in comparison to the M3 size standard marker. The marker alleles were scored as absent or present {0,1} and treated as dominant marker as described by Ghislain et al. (2009). Electrophoresis patterns of SSR markers are given in Appendix B (Figure B.1).

Cytoplasm type markers The QUEST population was genotyped with cytoplasm-specific markers to detect the cytoplasm type of each individual. The analysis was performed by Rena Sanetomo (NARO Hokkaido Agricultural Research Center, Japan). Six types of cytoplasms were distinguished as described in detail by Hosaka and Sanetomo (2012): T type (the most prevalent type in *S. tuberosum* spp. *tuberosum*), D type (introduced from *S. demissum*), A type (the most prevalent *S. tuberosum* ssp. *andigena* type), P type (introduced from *S. phureja*), and M type (Mother type, or an ancestral type of Andean cultivated potatoes) as well as W type (Wild species), including the *S. stoloniferum* derived sub-type.

Starch and yield associated PCR markers Six allele specific PCR markers and one epistatic interaction that were available for genes that are associated with starch and yield traits, were analyzed in the QUEST population. The markers and their associations are presented in Table 2.2. Primer sequences and annealing temperature are given in Appendix B (Table B.6).

Table 2.2.: **PCR markers associated with starch and yield in previous studies.** Arrows indicate the direction of the allele effect on the trait compared to the population mean

Locus	Chromosome	Marker allele	Trait	Source
Pho1a	3	*Pho1a-HA*	TSC ↑, TSY ↑	Schreiber et al. (*in preparation*)
Pain1	3	*Pain1-8c*	TSC ↑, TSY ↑	Li et al. (2008, 2013)
Pho1b	5	*StpL-3e*	TSC ↑, TY ↓	Li et al. (2008, 2013)
GP171	8	*GP171-a*	TSC ↓	Li et al. (2008)
HSP70	9	*HSP70-bad*	TSC ↓, TY ↑	Fischer et al. (*in preparation*)
Rca	10	*Rca-1a*	ns	Li et al. (2008)
		*Pain1-8c*Rca-1a*	TSC ↑, TSY ↑	Li et al. (2010)

Pho1a/Pho1b=starch phosphorylase 1a/1b, Pain1=potato vacuolar invertase 1, GP171=non coding genomic fragment, HSP70=heat shock protein 70, Rca=rubisco activase, TSC=tuber starch content, TY=tuber yield, TSY=tuber starch yield

Candidate gene genotyping SNP markers were developed for starch and yield associated candidate genes and scored in the QUEST population.

Gene coding sequence information was obtained from published accession numbers and retrieved from the NCBI database (NCBI; http://www.ncbi.nlm.nih.gov/). These sequences were then BLASTed against the potato genome sequence (version v4.03) (The Potato Genome Sequencing Consortium, 2011). Loci obtained, transcript or superscaffold numbers were then inserted in the PGSC Genome Browser (PGSC Genome Browser; http://solanaceae.plantbiology.msu.edu/cgi-bin/gbrowse/pgsc-potato-dm/) and loci were manually detected on the scaffolds, if not annotated. The position on the pseudomolecule, the genomic sequence as well as the exon-intron structure of the PGSC representative gene model were retrieved. When the locus was not annotated, the published cDNA sequences together with the reference genomic sequence were entered in NCBI Spidey (NCBI Spidey; http://www.ncbi.nlm.nih.gov/spidey/) to reveal the exon-intron structure of the gene as well as the position of the start and the stop codon.

Gene specific primers were designed to be located at the borders of the exons and to have a length of about 600 bp. Ideally a fragment included as little intron sequence as possible to avoid indels, that cause frame shifts in the sequencing output and make scoring

impossible. The primers used for amplicon sequencing are shown in Table 2.3.
The standard PCR reaction was performed in a 25 μl reaction volume, containing 50 ng genomic DNA, 10 mM Tris-HCL pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.1% Trifon X-100, 100 μM of each dNTP (Roth, Karlsruhe, Germany), 0.4 μM of each primer, 1U Ampliqon Taq Polymerase (Ampliqon, Odense M, Denmark) and deionized water (Merck KGaA, Darmstadt, Germany). Markers were amplified under the following PCR conditions: 3 min incubation time at 95°C, followed by 35 cycles of 94°C (20 sec), annealing temperature (Table 2.3) (40 sec) and elongation step at 72°C (60 sec), followed by 10 minutes final elongation at 72°C. The fragment length determined the elongation time at 72°C and was adjusted to 30 seconds per 500 base pairs. The amplification result was then checked on a 1.5% agarose gel.
Following PCR amplification, the products were purified with Illustra ExoStar (GE Healthcare Europe GmbH, München, Germany) for 15 min at 37°C, followed by 15 min at 80°C. PCR fragments were custom sequenced by the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de/home/) on Applied Biosystems (Weiterstadt, Germany) 3730XL Genetic Analyzer sequencer. Premixed reagents were purchased from Applied Biosystems and oligonucleotides for PCR as well as sequencing reactions were purchased from Sigma-Aldrich (Taufkirchen, Germany).
The generated sequences were aligned and SNPs were detected with the NovoSNP software (Weckx et al., 2005). The SNP dosages of the QUEST genotypes were then scored using both the Data Acquisition & Data analysis software DAx 8.1 (Van Mierlo Software Consultancy) and manual scoring. The SNPs were predominantly bi-allelic and were coded into five genotype classes {0,1,2,3,4}, with {0} being the class homozygous for the SNP allele represented by the potato genome sequence (AAAA), {1,2,3} the three heterozygous genotypes (AAAB, AABB, ABBB) and {4} the class homozygous for the second SNP allele (BBBB). SNPs with more than two alleles were excluded from the analysis.

2.2.6. Statistical analyses

Analyses were performed with the statistical software R (R Development Core Team, 2013) if not stated otherwise.

Phenotypic data analysis The phenotypic data of the two sites in 2010 and 2011 were compiled and treated to correct for the experimental design, using a linear model with two factors, implemented in the R package 'stats' using the function *lm()*. There was not sufficient data to correct for the experimental design at Neiker in both years, therefore correction was solely performed for the phenotypic data of Appacale. The factor *block* had 9 levels. The levels of the factor *tester* corresponded to the three tester varieties Desiree,

Table 2.3.: **Candidate gene loci that were sequenced and genotyped in the QUEST population.** Given are the primers, the annealing temperature for the PCR reaction, the fragment length as well as the sequencing primer to perform the Sanger amplicon sequencing for subsequent SNP scoring

Locus[1]	Chromosome	Primer	Primer sequence (5'-3')	T_a (°C)	length (bp)	scored SNPs/indels
CP12-2	1	CP12-2_4F[2]	GGCAACAATTGCTGGTGTTA	59	453	11/1
		CP12-2_3R	GCCTAATTCATAGCATTCAAGATTC			
SSsIV	2	SSsIVF1	CTCAATGAAGCTCGTGTCCA	50	869	15/–
		SSsIVR3[2]	CAAAATTCCGAAGGCATCTC			
PGM1	3	PGM1_1F[2]	ATGGCTATGGAGAGTGCATTGA	57	1060	17/–
		PGM1_1R	GTATCCAATTGGCAAGGTAATTGTC			
SssI	3	SssI-7_1F	GGATACTCATGGGAAATAACAACTCC	57	1022	18/1
		SssI-7_1R[2]	CAATCAGGTCGAATTGGAAGG			
PGI1	4	PGI1_1F	AGCATCTACTCACCTTCTTCATCTTTC	56	492	14/1
		PGI1_1R[2]	TGCAAACTGGCAAACAGCTT			
Pho1b	5	Pho1b_3F	TGTTGCAAGAAAAGCTAAACCAA	57	1178	23/1
		Pho1b_3R[2]	GATCACCAATCTCGGGATCA			
BEL5	6	BEL5_2F	CGATTATGGAAGCCAATGGT	57	668	20/–
		BEL5_2R[2]	GGAAATCGCTTATTCCCACTC			
BMY1	8	β-AmyI_F[2,3]	GCTACTGGAACATGGTGACAGA	60	342	10/–
		β-AmyI_R[3]	AGAGCATTTTCTCCAGCAAG			

[1] CP12-2=Chloroplast protein 12, SSsIV=soluble starch synthase 4, PGM1=phosphoglucomutase 1, SssI=soluble starch synthase 1, PGI1=phosphoglucoisomerase 1, Pho1b=starch phosphorylase 1b, BEL5=potato BEL1-like transcription factor 5, BMY1=beta-amylase 1; [2] Primer used for amplicon sequencing; [3] Primer from Krusiewicz et al. (2011)

Jaerla and Kennebec that were planted in each block. First the correction factors were estimated (Equation 2.4 and then the phenotypic data were corrected for the block effects (Equation 2.5).

$$y = \mu + \text{block} + \text{tester} + \varepsilon \tag{2.4}$$

$$y^* = y - \text{block} \tag{2.5}$$

Adjusted entry means over the two years and two sites were calculated from corrected data (Appacale) and primary data (Neiker) according to Li et al. (2008) using the linear model (Equation 2.6). Basis for the adjustment was the set of 50 tetraploid standard cultivars that were grown at both sites in both years. The levels of the factor *genotype* corresponded to the number of genotypes in the trial and factor levels for the factor *location* were 4: Appacale site in 2010, Appacale site in 2011, Neiker site in 2010 and Neiker site in 2011.

$$y = \mu + \text{genotype} + \text{location} + \varepsilon \tag{2.6}$$

Heritability is not only a measure for inheritance but if used for one generation with no selection, it gives an indication for the repeatability of the phenotypic data. It is calculated from the ratio between genotypic and phenotypic variance. Therefore, variance components were estimated by a mixed linear model with *location* as fixed and *genotype* as random term (Equation 2.7). The mixed linear model was implemented in the package 'lme4' by the function *lmer()*. Based on the estimated variance components, heritability (H^2) was calculated, where σ_g^2 represents the variance component for the genotypic main effect, σ_e^2 represents the variance component for the residuals and n the number of locations (Equation 2.8).

$$y = \mu + \underline{\text{genotype}} + \text{location} + \varepsilon \tag{2.7}$$

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{\text{n}}} \tag{2.8}$$

Partial correlations between phenotypic traits were calculated based on the adjusted entry means. Partial correlations describe the relationship between two traits without the effect of other traits. They were calculated for the 282 genotypes of the QUEST population with the function *cor()* (package 'corpcor'). A custom-made script was provided by Niklas Körber (MPIPZ, Cologne, Germany).

Assessment of hidden population structure Two approaches were applied for the detection of possible population structure: principle coordinate analysis and a Bayesian clustering approach.
The principle coordinate analysis (PCoA) (Gower, 1966) is based on pairwise genetic distances between the genotypes that were calculated from the microsatellite marker data that were sored as dominant markers. Jaccard's distances were calculated with R package 'prabclus' using the function *jaccard()* under default settings. Distances were transformed by square root transformation to obtain euclidean properties, following Reif et al. (2005). Principal coordinate analysis was performed based on Jaccard's distances between cultivars with the R package 'stats' by applying the function *cmdscale()*. The custom-made script was provided by Benjamin Stich (MPIPZ, Cologne, Germany). The explained variance of each principal coordinate was calculated. The total explained variance of the first eleven coordinates was 10% and the results of principal coordinates one to eleven were extracted for further analysis.
Population structure was further determined with a Bayesian approach, by analyzing the microsatellite marker data with the software STRUCTURE 2.3.4 (Pritchard et al., 2000). Burn-in time as well as iteration number was set to 100,000 with 10 repetitions, testing the probability of 20 subpopulations in the QUEST population. The results of the run were submitted to STRUCTURE HARVESTER (Earl and VonHoldt, 2012) and the most likely number of subpopulations was determined by the log likelihood combined with the Evanno method (Evanno et al., 2005).

Marker data and missing values Three markers with more than 5% of missing values were excluded from the analysis. Missing marker data was replaced at random, according to the proportion of genotypic classes within each individual marker. For SNP markers there were maximum five classes {0,1,2,3,4}, while missing values in SSR and indel markers were replaced with {0,1}, resembling the proportion of absent and presence. There were no missing values in the PCR marker data set and cytoplasm types. This method was tested in a subset of markers, where missing values were replaced three times at random. The difference between p-values of association mapping (see below) were neglectable.

Association mapping For association mapping, a two-step analysis approach was chosen, as described by Stich et al. (2008). First adjusted entry means were calculated and then association mapping was performed, using a mixed linear model, which accounts for population structure and kinship. The mixed linear model equation for the PK method (Stich et al., 2008; Yu et al., 2006) is shown in Equation 2.9. The population structure was accounted for by the P matrix, wherefore the first eleven principal coordinates were extracted, explaining 10% of the variance in sum. The K matrix represented the kinship

between genotypes as random effect. It was based on 183 dominantly scored SSR marker alleles using the 'EMMA' package (Kang et al., 2008).
In total, 309 markers were tested for associations with five phenotypic traits. 118 SNP markers and 4 indel markers from candidate genes, 181 SSR markers and 6 PCR markers. In addition, epistatic interaction between two PCR markers and association with cytoplasm-type was assessed. The analysis was performed with a mixed linear model (Zhang et al., 2010) implemented in the software package 'GAPIT' (Lipka2012). GAPIT was modified for the analysis of tetraploid data by Alexander E. Lipka (Cornell University, USA).

$$y = \mu + P + \underline{K} + \varepsilon \tag{2.9}$$

Linkage disequilibrium LD was estimated for all pairs of SNP markers with a chi-square test, based on an allele frequencies at all SNP loci. The obtained p-values were corrected with the Bonferroni-Holm correction (Holm, 1979) to account for the multiple testing problem. The analysis was performed using a custom-made script provided by Benjamin Stich (MPIPZ, Cologne, Germany).

2.3. Results

2.3.1. Phenotypic analysis

Phenotypic traits were assessed for 300 individuals that were planted at two trial sites in two subsequent years. The 50 wild *Solanum* species, with one plant per genotype grown from *in vitro* culture and cultivated in the tunnel, were not phenotyped.

Adjusted entry means Adjusted entry means were calculated for the 300 genotypes based on the experimental design (Appendix Table B.3). Boxplots of the adjusted entry means are shown in Figure 2.1. All groups of genotypes had a similar TSC and AMY, as the medians were similar for all groups. The medians of the different groups of genotypes for TY, TSY, TN and TW follow similar patterns, with the landraces having lower values compared to all other groups. TN in diploid clones was similar to tetraploids, while TW was much lower, showing that the lower yield of diploids was due to TW, not TN.
The histograms of the phenotypic trait values (Figure 2.2) all followed a normal distribution. The histogram of TN was slightly skewed to the left.

Heritability of phenotypic traits The heritability shows the repeatability of the phenotypic traits and was calculated from the ratio between genotypic and phenotypic variance (Table 2.4). TSC had the highest heritability (0.830), followed by TSY, TY, TN and TW. AMY had a very low heritability (0.028). This means that the variation in the trait is mainly due to environmental factors rather than to genetic factors. Therefore, AMY was excluded from association analysis.

Table 2.4.: **Estimated variance and heritability of phenotypic traits.** V_g represents the variance component for the factor genotype and V_e of the residuals. H^2 is the estimated heritability

Trait	TSC	TY	TSY	TN	TW	AMY
V_g	3.746	145,273	3,168.6	20.70	704.15	0.09
V_e	3.065	198,944	3,969.5	30.98	1,079.84	12.02
H^2	0.830	0.745	0.762	0.728	0.723	0.028

Correlations among traits Partial correlations where calculated between pairs of traits. They are a measure for the correlation between two traits with the effects of the other variables removed. In Figure 2.3, the partial correlations between the traits in the 282 genotypes of the QUEST population are shown by a correlogram.

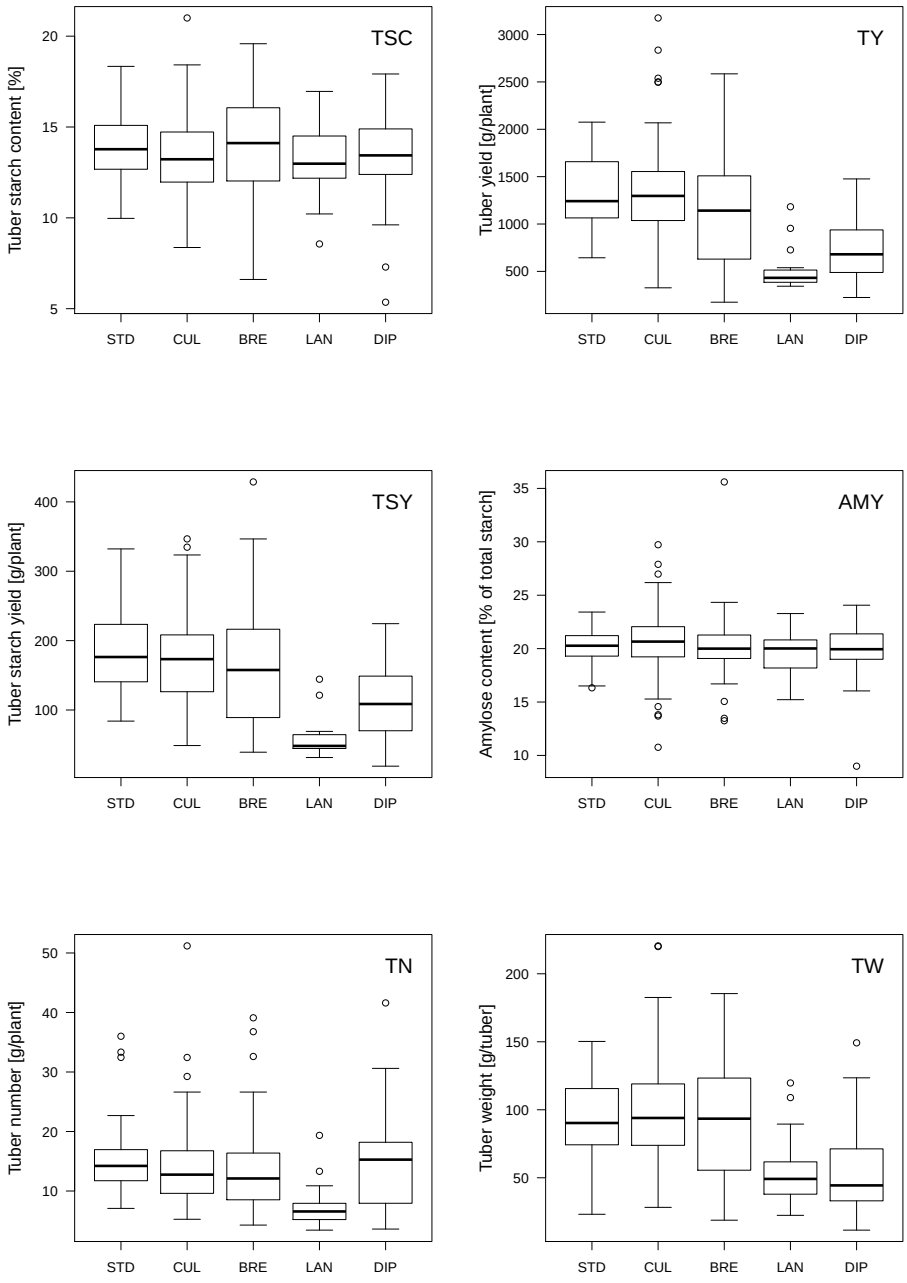


Figure 2.1.: **Box plots of the adjusted entry means.** Traits are depicted for tetraploid varieties [standard varieties (STD) and commercial cultivars (CUL)], breeding clones (BRE), landraces (LAN) and diploid clones (DIP)

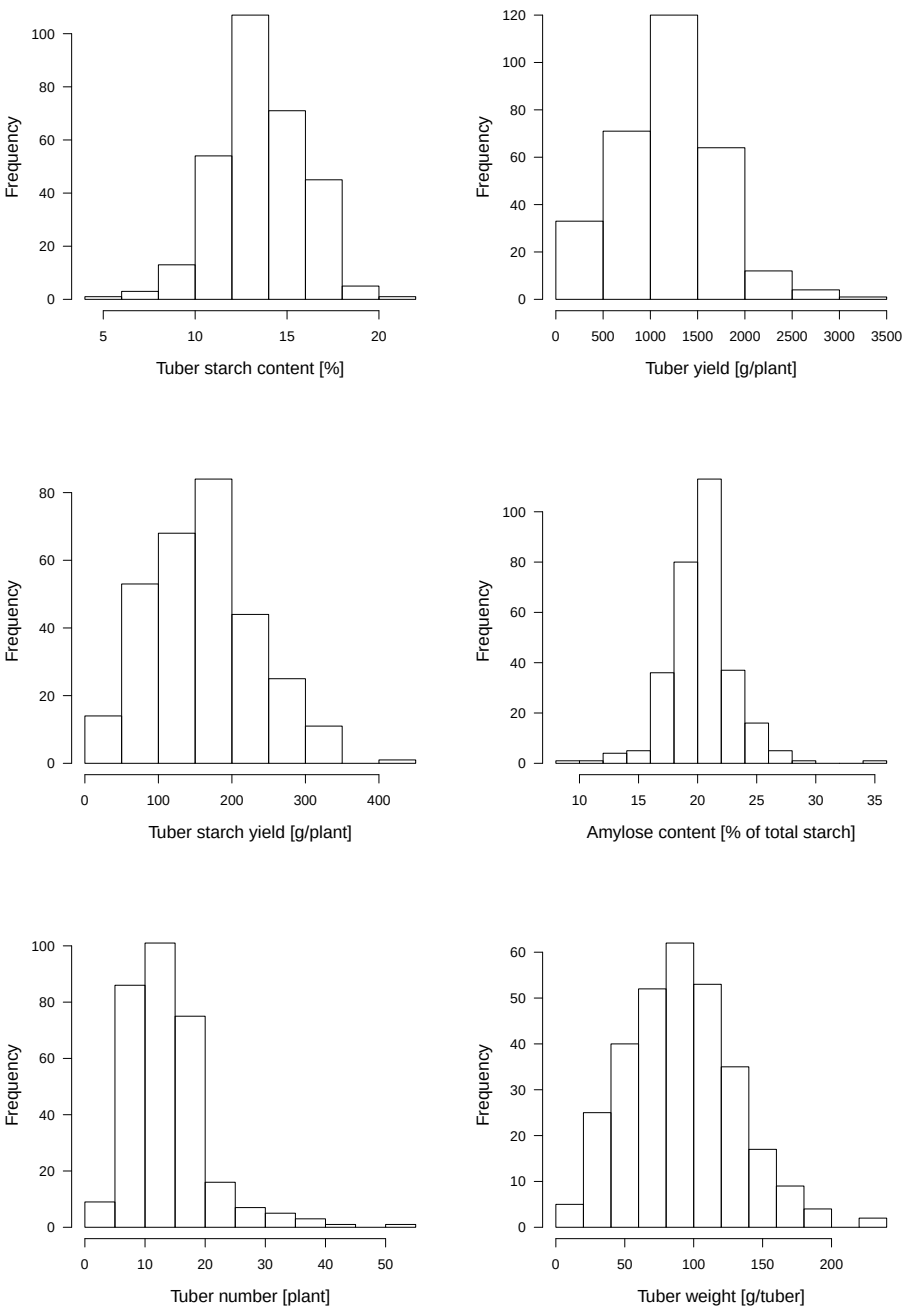


Figure 2.2.: **Histograms of the adjusted entry means.**

TSC and TY showed a strong negative correlation, which was highly significant (p-value <0.001). TSY shows a strong positive correlation (p-value <0.001) with both TY and TSC. Significant positive correlations existed between TY and TN as well as TW. A negative correlation was observed between TN and TW. The correlation coefficients and corresponding p-values are given in Appendix B (Table B.4).

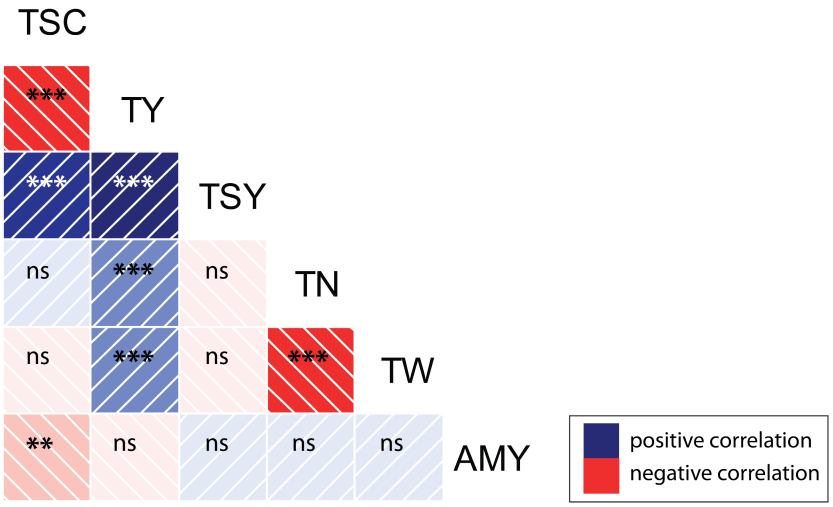


Figure 2.3.: **Partial correlations between pairs of traits for the genotypes of the Quest population** *significant at α=0.05; **significant at α=0.01; ***significant at α=0.001; ns=not significant.

2.3.2. Population structure analysis

282 individuals of the Quest population were genotyped with 29 microsatellite markers, including all varieties (standard and commercial cultivars), breeding clones and landraces. 183 alleles were sored for absence (0) or presence (1).

Principal coordinate analysis Based on the genotyping with microsatellite markers, Jaccard's distances were calculated and transformed to obtain euclidean properties (Reif et al., 2005). Figure 2.4 depicts a histogram of the estimated distances between genotypes. Principal coordinate analysis (PCoA) was performed, based on distance values. Figure 2.5 illustrates the 282 Quest genotypes, separated by Principal coordinate 1 (PC1) and Principal coordinate 2 (PC2). A large cluster of cultivars and breeding clones can be observed as well as a clustering trend of the landraces. In addition, a small cluster of four cultivars appears. The explained variance of PC1 was 1.18% and of PC2 1.08%.

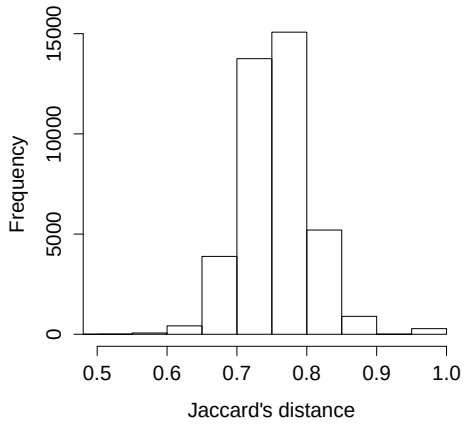


Figure 2.4.: **Distribution of Jaccard's distance estimates among genotype pairs of the Quest population**

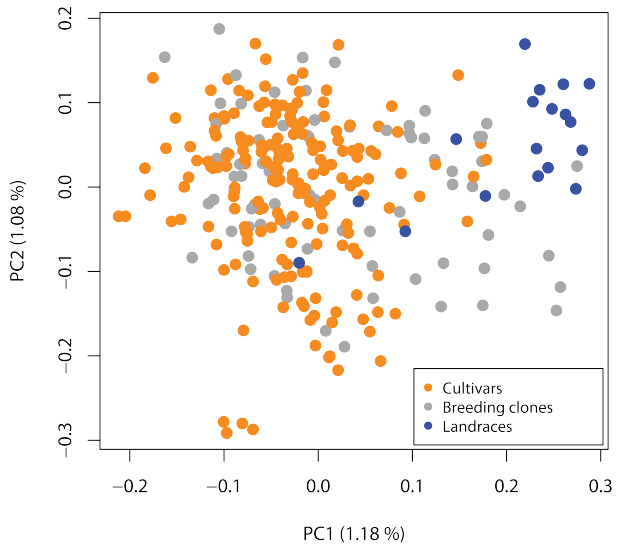


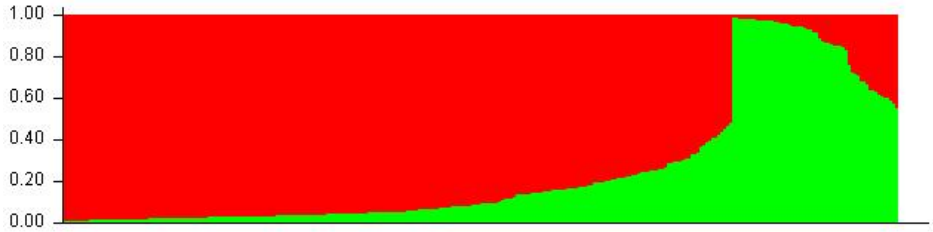
Figure 2.5.: **Principal coordinate plot of the Quest population.** The 282 genotypes of the Quest population were separated by the first two principal coordinates (PC) which were calculated on the basis of Jaccard's distances. Numbers in parentheses are the percentage of explained variance by the PC

Bayesian clustering The number of subpopulations in the QUEST population was furthermore assessed by Bayesian clustering, using the software STRUCTURE. The result of the analysis was processed with the STRUCTURE HARVESTER online tool (Earl and VonHoldt, 2012). The output of the log likelihood and Evanno method (Evanno et al., 2005) indicate the most likely amount of two subpopulations (K=2) in the QUEST population (see also Appendix A, Figure A.1). The barplots for K=2 are shown in Figure 2.6. When sorting the inferred subpopulations according to genotype group, the landraces show a more prominent representation of subpopulation 2. Representatives of subpopulation 2 can also be found in cultivars and breeding clones.

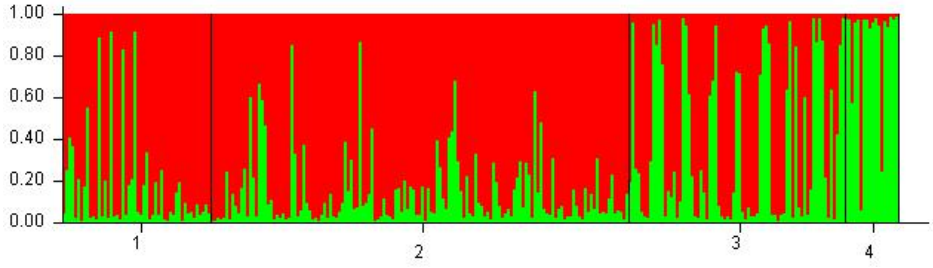
Cytoplasm type In order to find alternative underlying structure, cytoplasm types were determined. The 282 genotypes of the QUEST population were analyzed by Rena Sanetomo (NARO Hokkaido Agricultural Research Center, Japan) according to the method of Hosaka and Sanetomo (2012). Six different types of cytoplasm were present in the population. 144 of 189 cultivars showed T type cytoplasm (the most prevalent type in *S. tuberosum* ssp. *tuberosum*). D type (introduced from *S. demissum* and W type (Wild species) existed in about the same number in varieties (D=23 and W=22) and breeding clones (D=12 and W=15). 27 of 38 W cytoplasm types showed the W/γ sub-type of *S. stoloniferum*. A type (the most prevalent *S. tuberosum* spp. *andigena* type), P type (introduced from *S. phureja*) and M type (Mother type, or an ancestral type of Andean cultivated potatoes) were very rare (A=9, P=6 and M=1) and appeared only in landraces. The cytoplasm type was put into context with the number of inferred subpopulations from the Bayesian clustering approach. 90.6% of T type and 86.6% of W type was represented in subpopulation 1 (Table 2.5, Figure 2.6 c). A, M and P were solely represented in subpopulation 2, while D type was equally represented in both inferred subpopulations (52.8% and 47.2% respectively).

Table 2.5.: **Cytoplasm types in the K=2 subpopulations that were inferred by Bayesian clustering**

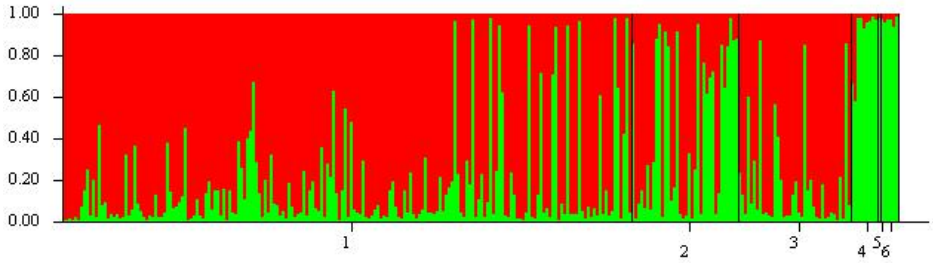
Cytoplasm type	Total	Subpopulation		% per subpopulation	
		1	2	1	2
T	192	174	18	90.6%	9.4%
D	36	19	17	52.8%	47.2%
W	38	33	5	86.6%	13.2%
A	9	0	9	0%	100%
P	6	0	6	0%	100%
M	1	0	1	0%	100%
Total	282	226	55		



(a) Barplot of inferred subpopulations K=2 ordered according to Q-value



(b) Barplot of inferred subpopulations K=2 ordered according to genotype group (tetraploid cultivars (1=standard varieties, 2=commercial cultivars), 3=breeding clones, 4=landraces)



(c) Barplot of inferred subpopulations K=2 ordered according to cytoplasm type (1=T type, 2=D type, 3=W type, 4=A type, 5=P type, 6=M type)

Figure 2.6.: STRUCTURE **graphical output of population structure for 282 genotypes at population size K=2, with genotypes ordered (a) according to the two subpopulations, (b) genotypes ordered to match the genotype group and (c) ordered according to cytoplasm type.** Individual genotypes are plotted on the x-axis. The probability (Q-value) of each genotype belonging to subpopulation 1 (red) or subpopulation 2 (green) is plotted on the y-axis

2.3.3. Candidate genes and mapping

Functional candidate genes related to starch and yield traits were compiled from the literature. Positions in the genome and gene copy numbers were detected by *in silico* mapping against the potato genome sequence (version v4.03). All functional candidate genes are presented in Table 2.6.
RFLP markers linked to starch and yield QTL (Schäfer-Pregl et al., 1998) were mapped to the potato genome sequence (version v4.03). The physical map of potato is shown in Figure 2.7 and includes also PCR markers as well as compiled candidate genes for starch and yield related traits. Chromosomes 7, 8 and 12 have a different chromosome orientation on the physical map when compared to the QTL map of Schäfer-Pregl et al. (1998). In a few cases the order of RFLP markers in QTL regions on the physical map was different from the mapping positions reported on the genetic maps. Some of the QTL from Schäfer-Pregl et al. (1998) fit to small regions of the physical map, other QTL spread over large regions of a chromosome, such as the QTL on chromosomes 1, 6 and 12 (Figure 2.7).

2.3.4. Association mapping

A total of 309 polymorphic markers that were scored in the 282 individuals of the Quest population were tested for associations the phenotypic traits TSC, TY, TSY, TN and TW using a mixed linear model including kinship and population structure. Kinship was assessed applying the 'EMMA' algorithm (Kang et al., 2008) and population structure was represented by the first 11 principal coordinates of the PCoA, explaining 10% of the variance. 118 markers from bi-allelic SNPs and 4 indel markers that were scored in amplicon sequences of eight candidate genes were tested for marker-trait associations as well as 181 microsatellite alleles from 29 loci and 6 PCR markers. In addition, epistatic interaction between two PCR markers and the cytoplasm type of the clones were tested for associations with the traits. The complete genotypic data is presented in Appendix B (Table B.7).
40 polymorphic markers (13%) were significantly (p-value <0.01) associated with at least one phenotypic trait. Marker-trait associations are given in Table 2.7 and Table 2.8. All markers associated with any trait (α=0.05) are reported in Appendix B (Table B.8). 32 of markers were associated with a single trait. 7 markers were associated with two traits. Two markers were associated with TSC as well as TSY, two markers with TY and TSY, one marker with TY and TW, one marker with both TSC and TN and one marker with TN and TW. 14 markers were associated with TSC, four markers with TY, eight markers with TSY, 14 markers with TN and seven markers with TW. The most significant associations were detected for TN, with three associations having p-values <0.001. One further highly significant (α=0.001) marker-trait association was detected for TSC.

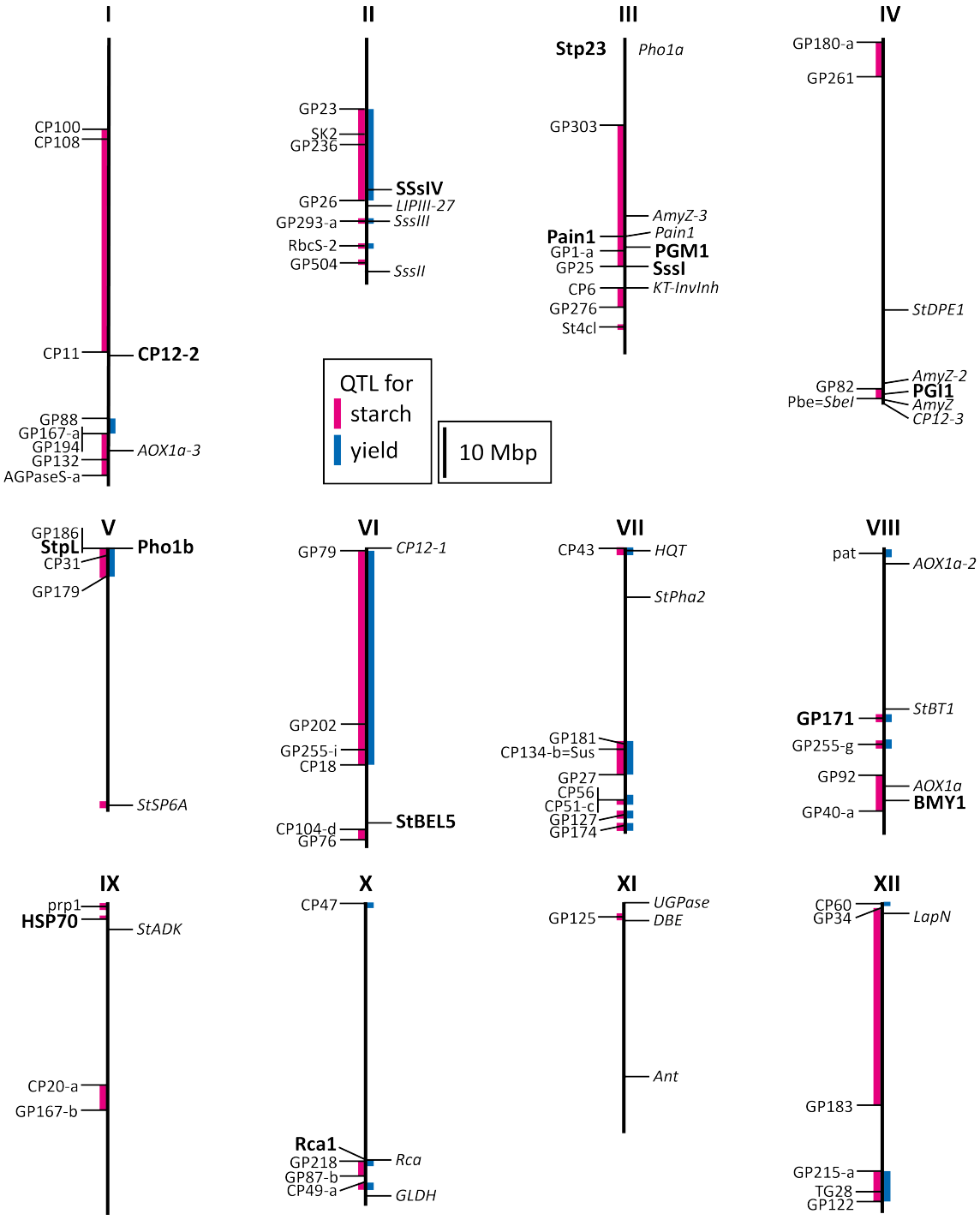


Figure 2.7.: **Physical map of potato (version v4.03) showing a selection of knowledge-based candidate genes for starch and yield traits.** Markers for tuber starch and yield QTL (Schäfer-Pregl et al., 1998) are indicated together with reported PCR markers associated with tuber starch content and tuber yield (Table 2.2). Markers and genes that were genotyped in this study are shown in bold

Table 2.6.: **Potato genes functionally related to tuber starch and yield traits.** Compiled information from literature search and *in silico* mapping (potato genome sequence, version v4.03)

Locus	Chromo-some	Gene number PGSC0003...	Position (bp) Start	End	Possible trait	Gene name; assumed gene function	Literature source
CP12-2	1	DMG400009042	62,723,718	62,724,360	yield	Calvin cycle protein CP12; Calvin cycle	Singh et al. (2008)
AOX1a-3	1	DMG400012558	81,916,714	81,920,957	starch, red.sugar	Alternative oxide synthase 1a	Krusiewicz et al. (2011)
SSsIV	2	DMG400008322	30,142,740	30,152,314	starch	Soluble starch synthase IV; starch granule formation in *Arabidopsis thaliana*	Roldán et al. (2007)
LIPIII-27	2	DMG400031758	33,058,256	33,059,792	starch	Triacylglycerol lipase III	Urbany et al. (2011)
SssIII	2	DMG400016481	36,377,154	36,379,312	starch	Soluble starch synthase III; starch synthesis	Abel et al. (1996)
SssII	2	DMG400001328	46,189,568	46,196,715	starch	Soluble starch synthase II; starch synthesis	Abel et al. (1996)
AmyZ-3	3	DMG400020603	35,999,186	36,002,230	starch	α-amylase; starch degradation	Li et al. (2008)
Pain1	3	DMG400013856	39,255,053	39,259,538	starch	Potato vacuolar invertase 1; starch-sugar interconversion	Li et al. (2008)
PGM1	3	not annotated	41,493,531	41,510,113	starch	Plastidial phosphoglucomutase; tuber starch synthesis	Tauberger et al. (2000)
SssI	3	DMG402018552	45,887,534	45,896,549	starch	Soluble starch synthase I; starch synthesis	Li et al. (2008)
KT-InvInh	3	DMG400010146	49,448,372	49,449,153	starch, yield	Kunitz-type Invertase inhibitor; invertase inhibition	Fischer et al. (2013)
Pho1a	3	DMG400007782	–	–	starch	Starch phosphorylase 1a;	Li et al. (2008)

Table 2.6.: (continued)

Locus	Chromo-some	Gene number PGSC0003...	Position (bp) Start	End	Possible trait	Gene name; assumed gene function	Literature source
						starch degradation, plastidial	
StDPE1	4	DMG400016589	53,993,893	54,001,156	starch	Disproportionating enzyme 1; starch degradation	Lloyd et al. (2004)
AmyZ-2	4	DMG400007974	68,256,932	68,260,197	starch	α-amylase; starch degradation	Li et al. (2008)
PGI1	4	DMG400012910	70,220,253	70,227,163	starch	Plastidial phosphoglucoisomerase; transitory starch breakdown in Arabidopsis leaves	Lu and Sharkey (2006)
AmyZ	4	DMG400009891	71,332,557	71,337,468	starch	α-amylase; starch degradation	Li et al. (2008)
CP12-3	4	DMG400009928	72,039,178	72,040,140	yield	Calvin cycle protein CP12; Calvin cycle	Singh et al. (2008)
Pho1b	5	DMG400028382	346,678	352,885	starch, yield	Starch phosphorylase 1b; starch degradation, plastidial	Li et al. (2008)
StSP6A	5	DMG400023365	51,319,128	51,320,774	yield	FT homolog; tuberization control	Navarro et al. (2011)
CP12-1	6	DMG400007286	195,064	195,709	yield	Calvin cycle protein CP12; Calvin cycle	Singh et al. (2008)
BEL5	6	DMG400005930	54,709,882	54,713,896	yield	Potato *BEL1*-like transcription factor; tuber formation	Chen et al. (2003)
HQT	7	DMG400011189	1,001,854	1,006,278	yield	Hydroxycinnamoyl CoA quinate transferase	Urbany et al. (2011)
StPha2	7	DMG400004101	10,832,288	10,839,563	starch	Plasma membrane H+-ATPase2 Driving proton coupled active sucrose transport	Li et al. (2008)

Table 2.6.: (continued)

Locus	Chromosome	Gene number PGSC0003...	Position (bp) Start	End	Possible trait	Gene name; assumed gene function	Literature source
AOX1a-2	8	DMG400018484	3,686,704	3,689,036	starch, red.sugar	Alternative oxide synthase 1a	Krusiewicz et al. (2011)
StBT1	8	DMG400020780	32,228,795	32,232,705	yield, starch	The Brittle1 protein homolog; plastidic adenine nucleotide uniporter	Leroch et al. (2005)
AOX1a	8	DMG400007614	47,496,732	47,500,233	starch, red.sugar	Alternative oxide synthase 1a	Krusiewicz et al. (2011)
BMY1	8	DMG400001855	50,592,099	50,595,681	starch, red.sugar	β-amylase 1; starch degradation	Scheidig et al. (2002)
StADK	9	DMG400027906	6,770,217	6,774,069	yield	Plastidial adenylate kinase; starch synthesis, tuber yield	Regierer et al. (2002)
Rca	10	DMG400019149	50,945,736	50,948,506	yield	Ribulose bisphosphate carboxylase activase; CO_2 fixation, Calvin cycle	Li et al. (2008, 2010)
GLDH	10	DMG400008132	58,305,327	58,313,142	yield	L-galactono-1,4-lactone dehydrogenase	Urbany et al. (2011)
UGPase	11	DMG401013333	808,268	814,810	starch	UDP-Glucose pyrophosphorylase; starch degradation	Sowokinos et al. (2004)
DBE	11	not annotated	3,945,875	3,946,053	starch	Debranching enzyme; starch synthesis	Kossmann et al. (1996) Chen et al. (2001)
Ant	11	DMG400013596	34,615,686	34,619,192	starch, yield	Adenylate transporter; starch synthesis	Chen et al. (2001)
StLapN	12	DMG400007831	2,328,178	2,335,525	starch, yield	Leucine aminopeptidase, neutral;	Fischer et al. (2013)

Direction of effects on a trait were different within loci. The three significant markers in the *SssI* for TW show opposite SNP allele effects. Also the two marker alleles of the *STI004* locus for TN show allele effects with opposing direction.

Marker-trait associations of polymorphic markers from candidate genes 14 biallelic SNP markers and one indel were significantly (α=0.01) associated with at least one trait of interest. These markers were scored at seven candidate gene loci, which encode the enzymes chloroplast-protein 12 (*CP12-2*), the soluble starch synthases I and IV (*SssI*, *SSsIV*), the plastidial phosphoglucomutase 1 (*PGM1*), the plastidial phosphoglucoisomerase 1 (*PGI1*), the starch phosphorylase 1b (*Pho1b*) and the BEL1-like transcription factor (*BEL5*). No association was detected at the β-amylase 1 (*BMY1*) locus.
The three SNPs and one indel in the *PGI1* locus were all associated with TSC. *PGI1_snp267* and *PGI1_snp252* had almost the same allele frequencies, p-values, R^2 and a negative minor allele effect. They were in high LD (q-value=1.2E-111) and therefore considered to be closely linked on the same haplotype. The presence of the *PGI1_indel202* minor allele had a negative effect on TSC and the *PGI1_snp333* minor allele has a positive effect on the trait. Two associations with tuber yield were identified in the two soluble starch synthases: *SssI_snp6015* was detected in a gene that has been described previously in potato, whereas the *SSsIV_snp2679* polymorphism was obtained from a novel candidate gene, which was found, based on the homology with the Arabidopsis *SSsIV* gene (Roldán et al., 2007). Two SNP markers, *PGM1_snp413* and *Pho1b_snp4319*, were associated with TSY. The strongest marker-trait associations (p-values <0.001) were found for TN with *CP12-2_snp327* and *BEL5_snp2960*. These SNPs had the highest amount of explained variance by the model, with 4.7% and 7.4%, respecively. Two further associations with TN were detected for *SssI_snp5995* and *Pho1b_snp4431*. Three SNP markers in *SssI* were significantly associated with TW.

Marker-trait associations of PCR markers and statistical epistatic interaction Six allele specific PCR markers (Li et al., 2008, 2013, Fischer et al., *in preparation*) and one statistical epistatic interaction (Li et al., 2010), for which associations with tuber and starch yield related traits were reported, were genotyped and tested for associations in the Quest population. In total, three marker-trait associations were detected. Two at significance level α=0.01 and one at significance level α=0.05.
Pain1-8c was positively associated with TSC and TSY, which confirmed the finding of Li et al. (2008). The presence of the *Rca-1a* allele was associated with TSC. There is a negative effect on the population mean in the presence of the allele in the Quest population. In the population of Li et al. (2008), this marker-trait association was not significant but showed a negative association with chips-quality. The *HSP70-bad* allele

Table 2.7.: **Marker-trait associations of candidate genes in the QUEST population.** All markers were tested with tuber starch content (TSC), tuber yield (TY), tuber starch yield (TSY), tuber number (TN) and tuber weight (TW)

Locus	SNP locus	SNP alleles	Minor SNP allele frequency	TSC p-value (R^2)	TY p-value (R^2)	TSY p-value (R^2)	TN p-value (R^2)	TW p-value (R^2)
Candidate gene markers								
CP12-2	snp327	T/C[1]	0.004 (C)[2]	ns[3]	ns	ns	0.000 (4.7) ↑[4]	ns
SSsIV	snp2679	A/T	0.062 (T)	ns	0.006 (2.4) ↑	ns	ns	ns
PGM1	snp413	A/T	0.035 (T)	ns	0.011 (2.1) ↓	0.008 (2.4) ↓	0.036 (1.5) ↓	ns
SssI	snp5871	A/G	0.446 (G)	ns	ns	ns	ns	0.009 (2.1) ↓
	snp5907	G/C	0.046 (C)	ns	ns	ns	0.011 (2.2) ↑	0.003 (2.8) ↓
	snp5995	C/T	0.008 (T)	ns	ns	ns	0.005 (2.7) ↑	ns
	snp6015	T/A	0.007 (A)	ns	0.005 (2.6) ↑	0.017 (1.9) ↑	ns	0.003 (2.7) ↑
PGI1	snp333	A/G	0.035 (G)	0.007 (2.3) ↑	ns	ns	ns	ns
	snp252(267)	T/A	0.082 (A)	0.006 (2.4) ↓	ns	ns	ns	ns
	indel202	wt/indel	0.076 (indel)	0.010 (2.1) ↑	ns	ns	ns	ns
Pho1b	snp4319	C/G	0.310 (G)	ns	0.014 (1.9) ↑	0.009 (2.3) ↑	ns	0.016 (1.8) ↑
	snp4431	A/T	0.006 (T)	ns	ns	ns	0.008 (2.5) ↓	ns
BEL5	snp2960	A/G	0.003 (G)	ns	ns	ns	0.000 (7.4) ↑	ns

[1]The nucleotide represented in the potato genome sequence stands on the first position; [2]Minor allele frequency of allele displayed in parenthesis; [3]Not significant at significance level α=0.01; [4]Arrows indicate the direction of the effect of the minor frequency allele on the trait compared to the population mean

Table 2.8.: **Marker-trait associations of PCR markers and SSR markers in the QUEST population.** All markers were tested with tuber starch content (TSC), tuber yield (TY), tuber starch yield (TSY), tuber number (TN) and tuber weight (TW)

Locus	Marker allele	Absent (0)/ present (1)	Present allele frequency	TSC p-value (R^2)	TY p-value (R^2)	TSY p-value (R^2)	TN p-value (R^2)	TW p-value (R^2)
PCR markers								
Pain1	Pain1-8c	0/1	0.170	0.000 (4.4) ↑	ns[1]	0.007 (2.5) ↑[2]	ns	ns
Rca	Rca-1a	0/1	0.443	0.008 (2.3) ↓	ns	ns	ns	ns
SSR markers								
STI043	STI043-e	0/1	0.504	0.005 (2.6) ↑	ns	0.005 (2.6) ↑	ns	ns
STI009	STI009-h	0/1	0.093	ns	ns	ns	0.009 (2.4) ↑	ns
STI001	STI001-b	0/1	0.404	0.029 (1.5) –	ns	ns	ns	0.003 (2.6) ↓
STI058	STI058-b	0/1	0.409	ns	ns	0.011 (2.2) –	ns	0.002 (3.0) ↑
	STI058-g	0/1	0.029	ns	ns	ns	0.001 (3.9) ↑	ns
STG0021	STG0021-a	0/1	0.039	ns	ns	ns	0.001 (3.6) ↑	ns
STM1043	STM1043-d	0/1	0.011	ns	0.043 (1.3) ↑	0.023 (1.8) ↑	0.000 (7.0) ↑	ns
STM1104	STM1104-f	0/1	0.125	0.002 (3.1) ↑	ns	ns	ns	ns
STG0025	STG0025-d	0/1	0.005	ns	0.041 (1.3) ↑	0.006 (2.6) ↑	ns	ns
STI028	STI028-f	0/1	0.317	ns	ns	ns	ns	0.001 (3.6) ↑
	STI028-b	0/1	0.493	0.002 (3.2) ↑	ns	ns	ns	ns
SSR20	SSR20-c	0/1	0.656	ns	0.017 (1.8) –	0.004 (2.9) ↓	ns	ns
STM1106	STM1106-b	0/1	0.122	0.001 (3.7) ↓	ns	0.032 (1.6) ↓	ns	ns
STI020	STI020-c	0/1	0.292	ns	0.024 (1.6) –	0.021 (1.8) ↑	0.006 (2.6) ↑	ns
STM3016	STM3016-f	0/1	0.297	0.003 (2.8) ↑	ns	ns	ns	0.043 (1.2) –
	STM3016-d	0/1	0.613	ns	ns	ns	0.010 (2.3) ↑	ns
STM0030	STM0030-e	0/1	0.507	ns	0.004 (2.7) ↑	0.003 (2.9) ↑	0.026 (1.7) ↑	ns
STI004	STI004-d	0/1	0.214	ns	0.037 (1.4) ↑	ns	0.002 (3.5) ↓	ns
	STI004-k	0/1	0.160	ns	ns	ns	0.005 (2.7) ↑	ns
STM3009	STM3009-b	0/1	0.007	0.002 (3.1) ↑	ns	ns	ns	ns

Table 2.8.: (continued)

Locus	Marker allele	Absent (0)/ present (1)	Present allele frequency	TSC p-value (R^2)	TY p-value (R^2)	TSY p-value (R^2)	TN p-value (R^2)	TW p-value (R^2)
STM1052	STM1052-c	0/1	0.712	ns	0.005 (2.5) ↑	0.002 (3.2) ↑	ns	ns
M17	M17-b	0/1	0.423	0.003 (2.9) ↑	ns	0.014 (2.0) ↑	0.023 (1.8) ↑	ns
M4	M4-a	0/1	0.979	0.003 (2.8) ↑	ns	0.018 (1.9) ↑	0.009 (2.3) ↑	ns
STM0003	STM0003-d	0/1	0.440	ns	ns	ns	0.006 (2.6) ↓	0.006 (2.3) ↑

[1]Not significant at significance level α=0.01; [2]Arrows indicate the direction of the effect of presence of the allele on the trait compared to the population mean

was associated with TY (α=0.05) in the QUEST population. The presence of the allele had a positive effect on the trait, compared to the population mean. This is in accordance with the findings of Fischer et al. (*in preparation*), who tested this allele for marker-trait associations in the CHIPS-ALL population.
There was no evidence for associations of *StpL-3e*, *GP171-a* and *Pho1a-HA* with any of the traits in the QUEST population. Furthermore, the statistical epistatic interaction between *Pain1-8c* and *Rca-1a* (Li et al., 2010) could not be detected, although it was very close to significance (p-value=0.054).

Cytoplasm type Cytoplasm type was not associated with the phenotypic traits in the QUEST population.

Marker-trait associations of SSR markers Of all 181 tested microsatellite marker alleles, 24 (13.8%) were significantly (α=0.01) associated with a minimum of one phenotypic trait. Five marker alleles were associated with two traits: one with TSC and TSY, two with TY and TSY, one with TSC and TN and one with TN and TW.
Eight marker alleles were associated with TSC, two with TY, five with TSY, 10 with TN and four with TW. The most significant (α=0.001) association was detected for TN. The SSR alleles *STM0037-a* and *STM0037-g* were associated with TSC in the study of Li et al. (2008). The marker-trait association of *STM0037-a* could not be confirmed in the QUEST population, but the *STM0037-g* allele was associated with TSC and TSY at significance level α=0.05. All SSR alleles that were significantly (p-value <0.01) associated with specific gravity in the study of Urbany et al. (2011) were included and tested in this study (*STI024-e*, *STI013-a*, *SSR327-a*). Specific gravity is considered equal to TSC. There was no evidence for the association of the *STI024-e* and *STI013-a* alleles with TSC in the QUEST population, but the *SSR327-a* allele was associated with TSC at significance level α=0.05. The *STM3009-b* allele, which was significantly associated with TSC, was present in only one individual. This individual showed the highest value of TSC in the QUEST population (Genotype N098).

2.3.5. Location of associated SNPs and amino acid exchanges

Four highly associated SNPs were found that lead to a non-synonymous amino acid exchange in the encoded protein. *SSsIV_snp2679* was located in exon 4 of the *SSsIV* gene (Table 2.9). The two different encoded amino acids were aspartic acid, encoded by the major frequency allele and valine, encoded by the minor frequency allele. Two SNPs at the *Pho1b* locus lead to a non-synonymous amino acid exchange. The minor frequency allele of *Pho1b_snp4319* lead to an amino acid change from proline to arginine. The *PGI1_indel202*

was the insertion of a triplet repeat AAG that lead to an extra lysine in the protein. The minor frequency allele of *Pho1b_snp4431* was causing an amino acid exchange from glutamic acid to aspartic acid. In *BEL5*, the minor frequency allele of *BEL5_snp2960* lead to a non-synonymous amino acid exchange from asparagine to serine.
The *CP12-2_snp327* was also located in an exon but causes a silent mutation. The same applied to the associated SNPs of *PGI1*. *PGI1_snp333*, *PGI1_snp252* and *PGI1_267* were located in an exon, but had synonymous effects. These mutations had no predicted effect on the protein composition (Table 2.9). The following SNPs were located in intron regions and did not encode any changes in proteins: *PGM1_snp413*, *SssI_snp5871*, *SssI_snp5907*, *SssI_snp5995* and *SssI_snp6015* (Table 2.9).

Table 2.9.: **Location of significantly associated SNPs in the gene and effects of SNP alleles on protein level**

Locus	SNP locus	SNP alleles	SNP type	amino acid exchange
CP12-2	snp327	T/C	synonymous	–
SSsIV	snp2679	A/T	non-synonymous	D <V
PGM1	snp413	A/T	intron	–
SssI	snp5907	G/C	intron	–
	snp5995	C/T	intron	–
	snp6015	T/A	intron	–
PGI1	snp333	A/G	synonymous	–
	snp252	T/A	synonymous	–
	snp267	G/A	synonymous	–
	indel202	insAAG	non-synonymous	extra K
Pho1b	snp4319	C/G	non-synonymous	P <R
	snp4431	A/T	non-synonymous	E <D
BEL5	snp2960	A/G	non-synonymous	N <S

2.3.6. Comparing mixed model and general linear model

In addition to the mixed linear model including population structure and kinship (MLM-Pk), all markers were analyzed with a general linear model (GLM) and a mixed linear model only accounting for kinship (MLM-k). This was done to get an idea how "robust" the significant markers are and how they perform in different models. Table 2.10 summarizes the output of the markers that were significantly associated with the traits in all three models. There were no significant marker-trait associations for TN in the MLM-k. The p-values of all models are given in Appendix B (Table B.9).

Table 2.10.: **Markers that were highly significantly associated in three different models.** Significance levels were α=0.001 in general linear model (GLM) and mixed linear model including kinship (MLM-k) and α=0.01 for mixed linear model including kinship and population structure (MLM-Pk). Markers with p-values <0.0001 in GLM and MLM-k are considered "robust" and are shown in bold

SNP locus	Model	TSC	TY	TSY	TN
CP12-2_snp327	GLM				**2.85E-06**
	MLM-k				**ns**
	MLM-Pk				**2.36E-04**
SssI_snp5995	GLM				5.54E-04
	MLM-k				ns
	MLM-Pk				0.005
PGI1_snp333	GLM	**1.83E-06**			
	MLM-k	**4.01E-06**			
	MLM-Pk	**0.007**			
Pho1b_snp4319	GLM			1.46E-04	
	MLM-k			2.09E-04	
	MLM-Pk			0.009	
BEL5_snp2960	GLM				**9.10E-09**
	MLM-k				**ns**
	MLM-Pk				**5.07E-06**
STI058-g	GLM				7.12E-04
	MLM-k				ns
	MLM-Pk				8.03E-04
STG0021-a	GLM				1.94E-04
	MLM-k				ns
	MLM-Pk				0.001
STM1043-d	GLM				**3.20E-05**
	MLM-k				**ns**
	MLM-Pk				**9.21E-06**
STM1104-f	GLM	**1.08E-05**			
	MLM-k	**1.72E-05**			
	MLM-Pk	**0.002**			
STM1052-c	GLM		4.37E-04	6.87E-04	
	MLM-k		5.67E-04	8.63E-04	
	MLM-Pk		0.005	0.002	
Pain1-8c	GLM	**2.79E-07**			
	MLM-k	**8.22E-07**			
	MLM-Pk	**2.32E-04**			
Rca-1a	GLM	**1.85E-05**			
	MLM-k	**3.37E-05**			
	MLM-Pk	**0.008**			

2.3.7. Linkage disequilibrium

LD between all pairs of SNP loci was tested with a chi-square test using the 118 bi-allelic SNP markers. q-values were obtained after correcting for multiple testing (Figure 2.8, Appendix Table B.10).

The total amount of marker pairs in significant (q-value <0.05) LD was 937 out of 6903 tested marker pairs (13.6%). Intralocus pairs in significant LD (α=0.05) in relation to the total amount of intralocus pairs were 32.8%. All interlocus pairs in significant LD in relation to the total amount of interlocus pairs were 10.9%. The highest LD was found between markers from same locus.

The candidate genes were located on chromosomes 1-6 and 8, with *PGM1* and *SssI* both located on chromosome 3. Of the 238 marker pairs tested on chromosome 3, 113 (47.5%) were in significant LD. Intralocus pairs of polymorphisms in significant LD were 33.2%, with half of the markers in *PGM1* and the other half in *SssI*. Interlocus pairs of polymorphisms in significant LD on chromosome 3 were 14.3%.

A pattern of LD, spanning five chromosomes was detected between SNPs at the *PGI1*locus (chromosome 4) with SNPs at the *SssI* (chromosome 3), *Pho1b* (chromosome 5), *BEL5* (chromosome 6) and *BMY1* (chromosome 8) loci.

Figure 2.8.: **Linkage disequilibrium between pairs of SNPs in candidate genes.** Displayed are -log q-values. Each row and line corresponds to one SNP position. Loci are framed with black boxes

2.4. Discussion

Candidate gene association mapping was applied to find natural variation at candidate gene loci with the aim to detect diagnostic SNP markers for tuber starch and yield-related traits. A novel association mapping population, the QUEST population, was assembled from varieties, advanced breeding lines and landraces. The genotypes were grown at two breeding stations in northern Spain in two subsequent years, with 50 standard varieties at both locations to facilitate joint statistical analysis of phenotype data. The population was phenotyped for tuber starch and yield-related traits: TSC, TY, TSY, TN, TW and AMY. AMY showed very low repeatability, suggesting a large environmental effect and was therefore excluded from association analysis. Marker-trait associations were detected using a mixed linear model, taking into account locations, years, population structure and kinship.

2.4.1. Novel marker-trait associations detected by candidate gene association mapping

Seven of the eight examined candidate gene loci show a minimum of one marker-trait association with at least one trait. This lead to a set of markers that can be directly applied in molecular marker-assisted selection in a breeding program. This result is not unexpected. The selection of candidate genes was knowledge-based. It focused on the functionality of the gene in the metabolic pathway of starch synthesis or its putative function for yield. Furthermore, genes in chromosomal regions that were co-localizing with QTL for starch content and yield (Schäfer-Pregl et al., 1998) were favored in the selection process. Therefore, the sequence variation of candidate genes might be directly associated with the observed phenotypic variation because it is located in the causal gene. However, association mapping makes use of the principle of LD, meaning that the phenotypic variation can also be caused by alleles of a gene that is in LD with the causal gene. Candidate gene association mapping has been shown to be a valid method for the detection of marker-trait associations in potato (Fischer et al., 2013; Gebhardt et al., 2004; Li et al., 2008, 2005; Malosetti et al., 2007; Pajerowska-Mukhtar et al., 2009; Simko et al., 2004; Urbany et al., 2011). The application of diagnostic SNP markers that were obtained by association mapping in potato breeding programs has been proven possible (Li et al., 2013). The diagnostic SNP marker that are detected by candidate gene association mapping can directly be implemented in breeding programs for potato starch yield optimization by marker-assisted selection although they might not be located in the causal gene.

Markers for tuber starch yield A significant marker-trait association for TSY was detected for the plastidial starch phosphorylase *Pho1b*. *Pho1b* plays a major role in the starch degradation and synthesis pathway. It was chosen as a candidate gene because of its co-localization with a QTL for tuber starch and yield (Schäfer-Pregl et al., 1998). Li et al. (2008) reported a significant marker-trait association of the *StpL-3e* SSCP marker allele with TSC and TY. However, there was no significant association of the *StpL-3e* allele with TSC or TY in the QUEST population. This could be due to the fact that the *Pho1b* amplicon was designed in 1,500 bp distance to the *StpL-3e* fragment and the LD was not high enough to detect the similar marker-trait associations. There was no association of the SNPs at the *Pho1b* locus with either TSC or TY, but *Pho1b_snp4319* was significantly associated with TSY. The minor allele had a positive effect on the trait and a frequency of about 30%. This frequency should be further increased by marker-assisted selection. Also, the SNP caused an amino acid exchange in the protein, which could potentially influence the protein structure and thereby its properties. In addition to the model accounting for population structure and kinship, *Pho1b_snp4319* was also highly significant in both GLM and MLM with only kinship. It can therefore be regarded as a rather "robust" marker. *Pho1b_snp4319* has a positive effect on TSY and can be of value in increasing desirable alleles in a potato breeding program for superior cultivars with improved TSY.
The plastidial *PGM1* generates glucose 1-phosphate, the substrate of AGPase and was therefore a functional candidate to yield markers for TSC and TSY. *PGM1_snp413* was associated with TSY, with the minor frequency allele having a negative effect on the trait. The SNP is located in the intron region of the gene, which indicates that it is physically linked or in LD with the causal SNP. The candidate gene was chosen based on the article of Tauberger et al. (2000), who conducted an antisense inhibition experiment of the plastidial phosphoglucomutase in potato. They report that potato lines with decreased activity of the plastidial *PGM* had a decreased amount of TSC, while tubers showed no morphological changes. Apart from the functional aspects, the gene was located in a QTL region for starch content (Schäfer-Pregl et al., 1998). Therefore marker-trait associations for starch content and starch yield were expected, provided that *PGM1* is the causal gene. One SNP marker for TSY could be detected in PGM1, which is suggested to be linked to the causal SNP and that can be directly applied to breeding for increased starch yield.

Markers for tuber yield Soluble starch synthase I (*SssI*) is a key enzyme in the metabolism of starch in potato tubers. In addition to that it is linked to in a starch QTL (Schäfer-Pregl et al., 1998). The *SssI_snp6015* polymorphism was positively associated with TY and the minor frequency allele is only present in simplex. The SNP is located in the intron region of the gene. Li et al. (2008) described an association with TSC and TSY.

Surprisingly, no marker-trait associations with either of these two traits were detected. It is possible, that the *SssI-4b* allele from the study of Li et al. (2008) was not detected as SNP by amplicon sequencing, although the PCR fragment was designed in a way that it included the complete amplified fragment of Li et al. (2008). The haplotype of the SSCP allele might be hidden within the detected SNPs and no haplotype tagging SNP might have been detected. A way of verifying this would be to test the amplicon for *SssI* in the genetic background of the Chips-All population. There were no marker-trait associations for TSC and TSY found in the *SssI* candidate gene, but one marker for TY could be detected.

The soluble starch synthase IV polymorphism *SSsIV_snp2679* is a potential marker for potatoes with higher TY. Three soluble starch synthases were described in potato (Werij et al., 2012). A fourth soluble starch synthase *SSsIV* was reported in *Arabidopsis thaliana* (Roldán et al., 2007). It was suggested to have a function in the control of starch granule formation in leaves. Furthermore, it was located in a QTL region for starch and yield (Schäfer-Pregl et al., 1998). Based both on function and position, *SSsIV* might be associated with TSC or TSY in potato. But instead, *SSsIV_snp2679* was associated with TY. A possible explanation for this result is that Arabidopsis has no storage organs like potato. Therefore, homologous genes can have a totally different function. Furthermore, the Arabidopsis *SSsIV* mutant shows not only a phenotype for increased starch granule size, but also a growth deficient phenotype. This indicates that the gene might have an additional impact on the general plant performance, which again could be influencing yield levels. Another explanation is, that the *SSsIV* candidate gene is not the causal gene. It might be physically linked or in LD with the causal gene and sequence variation of *SSsIV* is mirroring the sequence variation of the causal gene. The minor allele frequency was 0.062. This means that the allele was not very frequent in the tested plant material. As the effect of the minor frequency allele was positive, if can still be of interest for breeding. The *SSsIV_snp2679* polymorphism should be tested in other genetic background for verification of the marker-trait association with TY, but it is a good potential candidate for the optimization of TY by marker-assisted selection.

The *CP12-2* locus encodes a chloroplast protein 12 that is involved in the Calvin cycle (Singh et al., 2008), suggesting an influence on yield traits in potato. Three different *CP12* loci are present in the potato genome on chromosomes 1, 4 and 6. Kare Lehman Nielson from Aalborg University (Denmark) reported that a CP12 gene was found associated with potato TY in an RNA sequencing experiment (EAPR meeting, Wageningen, The Netherlands, 2010, oral presentation). Therefore a marker-trait association was expected with TY, which could not be detected in this study. There are several explanations for this. Firstly, the examined *CP12-2* on chromosome 1 might not be the corresponding

locus. Secondly, the relevant allele was not tagged by one of the SNPs. It was challenging to design an appropriate amplicon with sufficient polymorphisms for *CP12-2*, as the gene exists of one exon (data not shown) of approximately 400 bp. A third reason is that the relevant allele is either not present or very rare in the genetic background of the QUEST population and is therefore not associated. In association mapping, the statistical power to detect marker-trait associations of common variants is much higher than of rare alleles (Flint-Garcia et al., 2003). The best possible amplicon was designed for *CP12-2* and no association with TY was detected, but associations with TY might be found by testing further *CP12* loci.

Markers for tuber starch content *PGI1* is the plastidial isoform of phosphoglucoisomerase. It is active in transitory starch breakdown in the leaves of Arabidopsis (Lu and Sharkey, 2006). One indel and three SNP markers were associated with TSC, of which two SNPs were physically linked. *PGI1* is a novel candidate which has not been described in potato so far. For this study it was chosen based on functionality as well as its position in a starch QTL (Schäfer-Pregl et al., 1998). The significant SNP and indel markers were all located in the exon region of the gene, but all of them do not cause an amino acid exchange. *PGI1_snp333* was highly associated in all three tested statistical models and the minor frequency SNP allele had a positive effect on the trait. For these reasons it is a suitable marker for application in marker-assisted selection, as it is a "robust" marker. The *PGI_snp333* marker should be applied in marker-assisted selection for TSC in potato breeding programs, due to its high potential to be repeatable in other genetic background. The BMY1 locus encodes β-amylase I, a key enzyme in the degradation of starch in the plastid. However, no marker-trait association for TSC was detected in the QUEST population. Scheidig et al. (2002) showed the role of *BMY1* in transitory starch breakdown in leaves by antisense expression in potato plants. Expression data in the genome browser show, that it is not only expressed in leaves, but also in potato tubers. The candidate gene was selected based on its function but to the same extend because of its location in the region of a starch QTL (Schäfer-Pregl et al., 1998) on chromosome 8. It was expected to be associated at least with TSC. Provided that *BMY1* is a causal gene, there are several possible explanations for the missing marker-trait associations. One reason might be that the QTL in the study of Schäfer-Pregl et al. (1998) was detected based on an allele that is a rare allele in the QUEST population, as diploids seem rather different from tetraploids (Stich et al., 2013). Associations with rare alleles are unlikely to be significant in associations mapping, as the method is more suitable for the detection of associations with common variants (Flint-Garcia et al., 2003). Furthermore, it may be possible that the designed amplicon did not have sufficient SNPs to differentiate all haplotypes and the

SNPs are representing clusters of haplotypes. By this, the statistical power of a rare allele would be insufficient to allow a significant marker-trait association. A third explanation is that the allele of the QTL study (Schäfer-Pregl et al., 1998) is simply not represented in the QUEST population. As mentioned before, *BMY1* is involved in starch degradation process, possibly leading a higher amount of reducing sugars (Krusiewicz et al., 2011). Therefore it might play an important role in cold induced sweetening and could be directly linked to chipping quality in potato. The lack of significant marker-trait associations at the β-amylase I locus *BMY1* could be due to an under-representation of a rare allele in the QUEST population. Still, the fragment may be valuable for testing marker-trait associations with chipping quality in a different genetic background.

Markers for tuber number The marker-trait associations for TN might be interesting for marker-assisted selection of novel-type potato cultivars. The most significant marker trait associations at candidate loci are detected for TN, together with the largest amount of explained variance by the model. The minor alleles of *CP12-2_snp327*, *SssI_snp5995* and *BEL5_snp2960* had a positive effect on average TN. However, the genotype classes associated with higher TN are only represented by a small amount of individuals. Furthermore, the boxplots and histograms of TN illustrate the wide phenotypic variation in TN with up to 50 tubers in one plant in the case of commercial cultivars. This suggests that these cultivars have a large number of very small tubers. In agriculture practice, these mini-tubers are discarded directly on the field by the harvesting machine. Nevertheless, the market for fresh food potato varieties is open for new cultivars and varieties with small tubers might be of interest in the future. The *CP12-2_snp327*, *SssI_snp5995* and *BEL5_snp2960* polymorphisms are unlikely to be suitable for the breeding of starch potatoes, but could become interesting markers for breeding cultivars with an increased number of mini-tubers.

Markers for tuber weight Three marker-trait associations were detected for TW, all in the *SssI* locus, but they are unlikely to be suitable for marker-assisted selection. The minor frequency allele of *SssI_snp5907* had a negative effect on TW. Similarly, the minor frequency allele of of *SssI_snp5871* had a negative effect on TW, but the two SNP alleles were almost equally represented in the population. The minor allele frequency of *SssI_snp6015* showed a positive effect on TW, but the allele had an extremely low frequency. It is only present in a few genotypes in simplex and all SNPs are non-coding. Also, the polymorphisms are not highly significant in the other statistical models. This indicates that the polymorphisms are not sufficiently strong markers for the application in breeding programs. Therefore, the marker-trait associations for TW seem not suitable for direct implementation in a potato breeding program, unless the associations would be

confirmed in a different genetic background.

2.4.2. Known marker-trait associations verified in a novel genetic background

Allele specific associations from previous studies were validated in a novel genetic background confirming their value for marker-assisted selection. Three of six diagnostic SNP markers known to be associated with starch related traits in other populations showed similar associations in the QUEST population. The positive association of the *Pain1-8c* allele with TSC and TSY (Li et al., 2008) was confirmed. The *Rca-1a* allele was associated with TSC in the QUEST population. This result was unexpected, as the allele showed no significant association with TSC in previous studies. However, *Rca-1a* was associated with chips quality in the CHIPS-ALL population and the presence of the allele had a negative effect on the chips quality of chips. Bad chipping quality is caused by a higher amount of reducing sugars and therefore more degraded starch, resulting in a lower starch content. TSC and chips quality are positively correlated (Li et al., 2013; Werij et al., 2012), which means that a higher starch content is correlated with a higher chips quality. It can therefore be suggested, that the *Rca-1a* association with TSC was strong enough to show in the analysis in the QUEST population, opposed to the study of Li et al. (2008). *HSP70-bad* showed a moderate positive association with TY in the QUEST population. Fischer et al. (*in preparation*) reported that the presence of the allele showed a negative effect on TSC and a positve effect on TY in the CHIPS-ALL population (Li et al., 2008). Thus, one of the two associations could be confirmed in the novel genetic background. No marker trait associations were detected for *Pho1a-HA*, *StpL-3e* and *GP171-a*. Also the statistical epistatic interaction between *Pain1-8c* and *Rca-1a* was not detected in the QUEST population. Li et al. (2013) reported, that the marker-trait associations from previous studies can not always be verified in other populations. This may be due to the different genetic background and the changing environmental conditions of the trial years, influencing phenotypic data. The CHIPS-ALL population was evaluated in Northern Germany, whereas the QUEST population was evaluated in Northern Spain. The different latitudes could be causing genotype-by-environment interaction, leading to different associations. Several, though not all, reported marker-trait associations from previous studies for starch and yield traits in potato could be verified in the novel association mapping population.

2.4.3. Marker-trait associations detected for SSR markers

A number of new associations of SSR markers with traits were detected, while SSR marker-trait associations reported in previous studies could not be confirmed. 24 of 181 tested microsatellite alleles were associated with at least one phenotypic trait. The most interesting alleles, which were also highly significantly associated in the simple models, were

STM1104-f for TSC and *STM1052-c* for TSC as well as TY. *STM1104-f* is part of the granule-bound starch synthase I (GBSSI) encoding locus, which is active in the starch synthesis and highly influences starch composition in potato (Muth et al., 2008). These markers could have a potential for breeding programs. There is always a number of SSR markers associated with traits, but often these can not be verified in other studies. In this study, for example, all significant SSR marker-trait associations from Li et al. (2008) and Urbany et al. (2011) were included and tested in the QUEST population. Only two associations could be detected at a higher significance threshold (α=0.05) (*STM0037-g*, *SSR327-a*). As described above, this can be due to the fact, that the populations were grown at different latitudes and in different years, leading to genotype-by-environment interactions which influence the result of the association mapping.

2.4.4. Moderate population structure detected in commercially used potato germplasm

Population structure was analyzed in the QUEST population to avoid false-positive marker-trait associations in association mapping. Population structure was assessed with two approaches. With both the principal coordinate analysis and the Bayesian clustering method, the landraces were separated from the tetraploid cultivars and breeding clones. This could be due to the fact, that the landraces were varieties that are non-*tuberosum* species and had a different genetic background. Some also have other ploidy levels. In addition, the cytoplasm test showed that the landraces only had A, P and M type of cytoplasm and those three types were represented solely in subpopulation 2. This showed that the group of landraces were composed of germplasm different from *S. tuberosum*.
The detected population structure was only moderate. The clustering result of the PCoA had no strong statistical support as the proportion of explained variance compared to other studies (Stich et al., 2013) was rather low. Also, the visual output of STRUCTURE clarifies that the groups were not clearly separated from each other. These results prove the presence of a very moderate population substructure, which is supported by the findings of previous studies in tetraploid potato (D'hoop et al., 2010; Gebhardt et al., 2004; Li et al., 2008; Simko et al., 2004, 2006; Stich et al., 2013; Urbany et al., 2011). These findings are in contrast to Hamilton et al. (2011), who described a quite distinct population substructure. However, they were testing a much more diverse set, including diploid breeding lines, genetic stocks and cultivated *Solanum* species next to commercial tetraploid potato cultivars.
The data suggest an underlying selection for cytoplasm types by breeding programs. The output of the Bayesian clustering method visualized that the frequency of genotypes present in subpopulation 2 (green) tended to decrease from breeding clones to varieties.

So those results might show that subpopulation 2 is slowly being eliminated through the breeding process. T, D and W type of cytoplasm preliminary existed in the varieties and breeding clones. Also almost all T type and W type were represented in subpopulation 1 (red). In contrast D type was equally represented in both subpopulations. Therefore, those extreme differences in frequency might suggest that the T type and W type were preferentially selected in subpopulation 1 as opposed to D type and other types.

2.4.5. Linkage disequilibrium between alleles of unlinked loci

There was evidence of LD between unlinked markers that can lead to the detection of marker-trait associations located in distant genomic regions. More than 10% of all inter-locus SNP pairs - interchromosomal as well as intrachromosomal - were in significant LD, which was higher than expected after the findings of Pajerowska-Mukhtar et al. (2009). LD between unlinked loci can have several reasons, such as selection, mutation, mating system, population structure (Soto-Cerda and Cloutier, 2012). Basically, association mapping relies on the mechanism of LD. Marker-trait associations can be detected because quantitative trait loci are in linkage disequilibrium with a polymorphic marker. LD can be broken by recombination events, so the best possible marker lies directly in the causal gene. In potato there are large haplotype blocks, which are due to a rather low number of meiotic events between genotypes (Gebhardt et al., 2004). This favors the detection of marker-trait associations of markers that are not located in the causal gene. Still, the detected markers can be applied directly in breeding programs. The disadvantage of large haplotype blocks can be that it is hard to put hands on a causal gene. The causal gene and the molecular marker-trait association can be distal to each other due to strong LD between unlinked loci. However, the associations found by candidate gene association mapping can still be implemented in potato breeding programs.

The majority of marker pairs in LD (32.8%) were found within loci, which is due to physical linkage. Marker pairs of the 118 bi-allelic SNP markers that were detected by candidate gene amplicon sequencing were tested for LD with a chi-square test. More than thirty percent of the intralocus pairs showed significant LD and the highest LD was found between markers from same locus. This result was expected, due to the physical linkage of SNP alleles within the same locus. Physical linkage leads to increased LD values, meaning that there is a non-random association of alleles in the population (Soto-Cerda and Cloutier, 2012). This result is comparable to the findings of Pajerowska-Mukhtar et al. (2009), who found that almost half of the detected polymorphisms in LD were intralocus pairs.

The special pattern of LD between SNP alleles spanning a number of chromosomes might be due to the co-selection of haplotypes. The LD plot showed a pattern of SNP alleles in

LD, spread over five chromosomes. Alle genes (*PGI1, SssI, Pho1b, BEL5, BMY1*) were candidates that function in a joint network for starch synthesis and degradation. It may be that certain alleles are under co-selection with each other. The would explain why those SNP alleles occur together more frequent than expected under the assumption that they are independent from each other. Therefore, the possible co-selection of haplotypes of candidate genes within a network may be detected by the LD between SNP alleles, which spread over five chromosomes.

2.4.6. Limitations of candidate gene association mapping

Candidate gene association mapping is a knowledge-based method for the detection of diagnostic SNP markers. The method makes use of information, that is present in the literature. Appropriate candidates are chosen based on their functionality in pathways or co-localization with QTL regions from mapping studies. The only limitation is the researcher's knowledge about pathways as well as creativity, when it comes to elaborate on putative function of genes from other organisms. In order to find really novel diagnostic markers, an unbiased approach may be more suitable.
Linkage disequilibrium is a limiting factor in the detection of diagnostic SNP markers by candidate gene association mapping. The larger the physical distance between the marker and the causal gene, the higher the risk that the LD between marker and causal gene is broken by recombination during breeding. The ideal marker for a trait lies therefore directly in the candidate gene. For finding markers directly in causal genes, a low LD would be favorable, although the required marker density would rise drastically. Until now, there are several different claims on the extend of LD in potato as well as the necessary marker density for genome-wide association studies, ranging from 275 bp (Stich et al., 2013) up to 70 kb Simko et al. (2006). However, if LD really decayed as rapidly as reported by Stich et al. (2013), this would mean that almost all associations detected in the candidate gene association mapping approach would be directly located in causal genes. The large number of marker-trait associations that has been found so far, with a limited number of markers indicate that the extend of LD must be larger than 275 bp.

2.4.7. Concluding paragraph

The results from this study demonstrate that diagnostic SNP markers for the optimization of starch yield by marker-assisted selection can be identified by candidate gene association mapping. It has proven to be a valuable tool for the detection of diagnostic molecular markers for traits that can only be assessed late in the breeding process. Marker-assisted selection for these traits speed up the breeding process, saving valuable resources and time.

3. Potato tuber starch and yield markers identified by SolCAP Potato Array genotyping in a case-control design and association mapping

3.1. Background

Potato (*Solanum tuberosum*) is a tetraploid and highly heterozygous crop which is of major importance for the food, the feed and the industrial use. The market share of starch potatoes is growing (Statistisches Bundesamt, 2012). Breeding for special traits by molecular marker-assisted selection is the focus of potato breeders. However, there is a limited number of molecular markers available for the desired traits.
Higher throughput genotyping assays, such as the SolCAP Potato Array, have been recently developed (Felcher et al., 2012). High-density genotyping of potato germplasm with 8,303 SNPs can be performed at relatively low cost. In a case-control study, two groups of unrelated genotypes with extremely high and low values for a trait are screened by genotyping. Information about the genetic background of the differences is obtained by comparing the groups (Balding, 2006). The combination of a case-control design with high-density genotyping could be a valuable tool for the detection of novel markers.

Approach The major goal of this study was the discovery of novel candidate genes for tuber starch content, tuber yield and tuber starch yield. Therefore, three case-control studies were designed. Selected potato cultivars and breeding clones, were assembled in pools that showed extremely high and low values for the desired traits. About 48 genotypes, per case-control study, were assigned to the two pools and all individuals were genotyped for 8,303 SNPs with the SolCAP Potato Array. SolCAP SNPs significant for tuber starch and yield-related traits were selected from the results of the case-control studies and validated by association mapping in the full Quest population.

3.2. Materials and Methods

3.2.1. Case-control design

For this study, 90 genotypes were selected, based on the adjusted entry means, from tetraploid cultivars and breeding clones of the QUEST population (Chapter 2). 48 genotypes with extremely high and extremely low values of tuber starch content (TSC) and tuber yield (TY), respectively, were assigned to two case-control populations. A third case-control population for tuber starch yield (TSY) was established from 45 genotypes (Table 3.1, Table 3.2). The phenotypic data of the case-control pools are depicted in Figure 3.1.

Table 3.1.: **Genotypes in the case-control studies.** The genotypes were selected from the QUEST population, based on the adjusted entry means for tuber starch content (TSC), tuber yield (TY) and tuber starch yield (TSY). 'high' and 'low' represent the deviation of the extreme phenotypes from the population mean

Trait	'high' pool	'low' pool	Total
TSC	24	24	48
TY	24	24	48
TSY	21	24	45

Table 3.2.: **Genotypes selected from the QUEST population for the case-control studies.** Indicated are the number of genotypes grown at the two breeding stations Appacale and Neiker

	Cultivars	Breeding clones	Total
Appacale + Neiker (Standards)	9	–	9
Appacale	15	34	49
Neiker	24	8	32
Total	48	42	90

3.2.2. Sample preparation and SNP genotyping

Genomic DNA was extracted as described in Chapter 2. After extraction, the DNA samples were diluted to 200 ng genomic DNA in 4 μl. The 90 genotypes together with 6 technical replicates were genotyped individually with the Infinium 8,303 Potato Array at the Life & Brain Center (Department of Genomics, Bonn, Germany) on an Illumina iScan system, applying the Infinium assay.

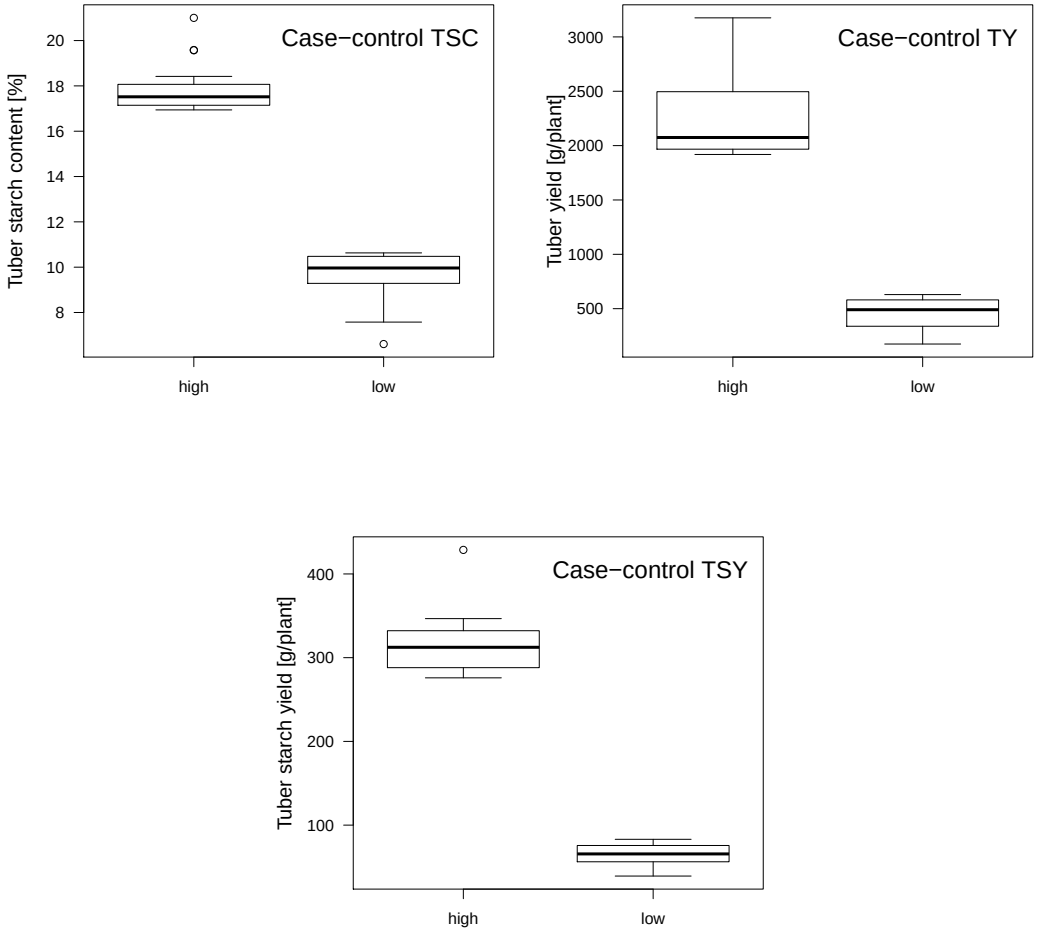


Figure 3.1.: **Box plots of the adjusted entry means of case-control populations**

The genotypes of the SNP markers were called from the bi-allelic marker data of the tetraploid potato clones, implementing the software package 'fitTetra' (Voorrips et al., 2011). The package was run using the function *fitTetra* with the option *try.HW=F*. The service company provided the script for the data analysis, which was applied by Shyamkumar Immadi. Each individual was assigned to one of the five possible genotypes at one SNP marker locus.

3.2.3. Statistical analysis of case-control studies

Pearson's goodness-of-fit test (chi-square test) was applied to test the null hypothesis that the allele frequency of the bi-allelic SNP makers in the contrasting pools are equal. The null hypothesis was rejected at the significance level α=0.01. This means that with p-values lower than 0.01, it was accepted that the allele frequencies in the two pools were significantly different. The chi-square tests were performed by Enrique Ritter (Neiker, Vitoria-Gasteiz, Spain) using the software SAS (version 9.1).
Six pairs of technical replicates were included in the analyzed genotypes. For each pair the Person's correlation coefficient was calculated to assess repeatability of the method.

3.2.4. Candidate loci selection

In the case-control studies, only genotypes with the most extreme values for the trait of interest were included. In the full Quest population, these traits had a quantitative distribution (Chapter 2). A subset of six SolCAP SNPs at relevant candidate loci was selected from the SolCAP marker loci to validate these marker-trait associations in a broader genetic background. Three major selection criteria were applied: significance, genomic position and function. In detail, all SolCAP SNPs with significantly different (α=0.01) allele frequency between the contrasting pools were mapped to the potato genome sequence (version v4.03) for each case-control study individually. Positions showing a cluster of significant SNPs were considered more interesting than positions with isolated SNPs. In a clustering region, the SNP with the strongest association was selected and the annotated genes in the area of about 500 kb around the SolCAP SNP was investigated to detect potential candidate genes for the trait of interest. The main focus lay on regions with little previous marker information. Major starch and yield loci in QTL regions, such as the *AGPaseS-a* locus on chromosome 1 (Schäfer-Pregl et al., 1998), were excluded. The function of the gene including the SolCAP SNP was the least important selection criterion, as it was considered unlikely to find the causal gene for a trait. However, genes in an interesting region, that were reported to act in carbohydrate or yield related pathways and contained highly significant SolCAP SNPs were favored. The selected, novel candidate loci with their associated trait and putative function are described in Table 3.3.

Table 3.3.: **Selected candidates based on statistical output of case-control studies.**

Locus	SolCAP id (solcap_snp...)	Chromosome	Position (bp)[1]	Significant trait	PGSC gene id (PGSC0003...)	Annotation in the genome browser; Putative function/metabolic role
AP2TF	c2_16349	2	14,356,715	TSC, TY, TSY	not annotated	AP2-like (ethylene-responsive) transcription factor At2g41710
	c2_16350		14,356,716	ns[2]		
F2PA	c2_11924	5	3,708,760	TSC	DMG400030565	Fructose-bisphosphate aldolase; sequence homology to *S. tuberosum* plastidial aldolase (NM_001288043), functional in Calvin cycle
	c2_11925		3,708,757	ns		
CP12-1	c2_54011	6	195,511	TY	DMG400007286	Chloroplast protein 12; Calvin cycle and non-photosynthetic tissues
QUA1	c2_9204	6	58,259,275	TY	DMG400020103	Glycosyltransferase QUASIMODO1
	c2_9203		58,259,535	ns		
60S	c2_3063	9	58,375,217	TY, TSY	DMG400029622	60S acidic ribosomal protein PO; Ribosomal subunit, located in cytoplasm
CIS	c2_25372	12	1,324,476	TSC, TY, TSY	DMG400007797	Citrate synthase; first enzyme of the citrate acid cycle

[1] According to potato genome sequence version v4.03; [2] ns=not significant at significance level α =0.01

3.2.5. Genotyping for association mapping

Two methods were chosen for genotyping: pyrosequencing and amplicon sequencing. The methods are described below. The pyrosequencing assay allowed for lower-cost genotyping of individual SNPs. Candidate genes that had an intron-exon structure which enabled the design of longer PCR fragments were genotyped by amplicon sequencing. This resulted in additional SNP loci in those candidate genes. The primers and PCR conditions for the pyrosequencing assay as well as for amplicon sequencing are given in Table 3.4.

Amplicon sequencing of candidate genes Primer design, amplicon sequencing and SNP detection were performed as described in Chapter 2. In brief, a PCR product of about 500 bp length was amplified. The fragments were sequenced with Sanger sequencing at the Max Planck-Genome-centre Cologne (Germany) on an Applied Biosystems (Weiterstadt, Germany) 3730XL Genetic Analyzer sequencer. Polymorphisms were scored with the Data Acquisition & Data analysis software DAx 8.1 (Van Mierlo Software Consultancy) as well as manual scoring.

Pyrosequencing for genotyping of SNPs The pyrosequencing assay is based on sequencing by synthesis (Ronaghi et al., 1998). The method has successfully been applied for the genotyping of single nucleotide polymorphisms in tetraploid potatoes (e.g. Draffehn et al., 2010; Rickert et al., 2002).
PCR primers were designed as described for Sanger sequencing, but with a fragment length of 100-300 bp in such a way that the significant SolCAP SNP was targeted. The sequencing primer was designed to anneal to the template with a maximum distance of 3 bp between the 3' end of the primer and the SolCAP SNP. The PCR primer that annealed on the strand opposite to the sequencing primer was biotinylated on the 5' end.
PCR fragments for pyrosequencing were amplified in a 25 μl reaction volume, containing 50 ng genomic DNA, 10 mM Tris-HCL pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 0.1% Trifon X-100, 200 μM of each dNTP (Roth, Karlsruhe, Germany), 0.4 μM of each primer, 1U Ampliqon Taq Polymerase (Ampliqon, Odense M, Denmark) and deionized water (Merck KGaA, Darmstadt, Germany). PCR fragments for pyrosequencing were amplified under following PCR conditions: 2 min at 94°C, followed by 50 cycles of 93°C (45 sec), annealing temperature (45 sec) and 72°C (60 sec), completed by a final elongation step of 10 min at 72°C. The success of the PCR and the intensity of the bands was assessed on a 1.5% agarose gel.
PCR products were prepared for pyrosequencing and the pyrosequencing was performed, following the protocol given in Appendix B (Protocol B.1). The samples were sequenced with the Pyrosequencing PSQ96 MA System (Biotage AB, Uppsala, Sweden) using the

Pyromark Gold Q96 reagents kit (Qiagen, Hilden, Germany). Signals were analyzed, using the software supplied by the manufacturer.

3.2.6. Association mapping

The SNPs from the selected candidate loci obtained by genotyping of the QUEST population were tested for marker-trait associations, implementing the R package 'GAPIT' (Lipka et al., 2012). Association mapping was performed with a mixed linear model, including population structure and kinship. Data preparation and statistical analyses are specified in Chapter 2.

3.2.7. Analysis of linkage disequilibrium

Linkage disequilibrium (LD) between significant SolCAP SNP loci from the case-control studies was assessed by a chi-square test like described in Chapter 2. LD was also tested for candidate genes, that were selected based on the output of the case-control studies.

Table 3.4.: **Selected candidates gene loci that were genotyped in the** QUEST **population.** Genotyping was either performed by amplicon sequencing or pyrosequencing. A biotin label at the 5'end of a pyrosequencing primer is indicated by '(Bio)'

Locus[1]	Chromosome	Primer	Primer sequence (5'-3')	T_a (°C)	length (bp)	scored SNPs
AP2TF	2	AP2TF_4F[2]	GCTCGTCAAACAGATTCCCATC	57	179	4
		AP2TF_3R	GCTCATAGCAGACAATGTTGAGCTT			
F2PA	5	F2PA_2F-B	(Bio)CCATTGATGAATCGAATGCAA	57	181	2
		F2PA_3R	TTCTTCCCATCGGTAGTGGAC			
		F2PA_2S[3]	TGTGTTGTCCAGACCAATTGA			
CP12-1	6	CP12-1_4F-B	(Bio)CCGATTCCCCTAAAATTCAACC	56	147	1
		CP12-1_5R	CACTTTCCGCAACTAAAATCAGA			
		CP12-1_4S[3]	CATACATCCTCCCGTAA			
QUA1	6	QUA1_4F	GAGCTTGATTGCTGCCAAGT	56	779	9
		QUA1_3R[2]	ATTCCTCCGTGAGCTTCTCC			
60S	9	60S_3F[3]	GCAAGGGATTCGTAAGGGTCT	56	107	1
		60S_3R-B	(Bio)CATCGTTTCCGGTACTTTTCG			
CIS	12	CIS_5F[2]	GGATGCGTCACTATATGCCTCT	56	511	7
		CIS_5R	CCTGGGATGCATGCTACAAT			

[1] AP2TF=AP2-like transcription factor, F2PA=Fructose-bisphosphate aldolase, CP12-1=Chloroplast protein 12, QUA1=QUASIMODO1, 60S=60S acidic ribosomal protein PO, CIS=Citrate synthase; [2] Sequencing primer in Sanger amplicon sequencing; [3] Sequencing primer in pyrosequencing assay

3.3. Results

3.3.1. SolCAP SNP genotype calling

90 potato cultivars and breeding clones were analyzed for 8,303 SNP loci with the SolCAP Potato Array. The bi-allelic marker data were analyzed using the R-package 'fitTetra' (Voorrips et al., 2011). 2,238 SNPs were excluded in the scope of the analysis and genotypes were assigned to 6,065 SNPs, corresponding to a success rate of 73%.
Additionally, six technical replicates were genotyped to observe the repeatability of the method. The Person's correlation coefficient was calculated for each pair of technical replicate, based on the 6,065 high quality polymorphic SolCAP SNPs. Correlations between technical replicate pairs ranged between 94-97% (Appendix Table B.13). Because of the high correlation between pairs, only one replicate of each pair was kept for the further analysis.

3.3.2. Statistical analysis of case-control studies

The 90 genotypes were assigned to the designated 'high' and 'low' pools of the case-control studies. The case-control studies were analyzed with a chi-square test for different allele distribution between contrasting pools. In total, 328 SolCAP SNPs had significantly different allele frequencies between pools in at least one of the case-controls studies (Appendix Table B.17). 207, 85 and 76 significant SolCAP SNPs were obtained from the case-control studies for TSC (Appendix Table B.14), TY (Appendix Table B.15) and TSY (Appendix Table B.16), respectively. The results of the statistical analysis of the case-control studies are summarized in Figure 3.2. The two SolCAP SNPs that were significant in all three case-control studies were *solcap_snp_c2_16349*, that mapped to a region with no annotation on chromosome 2, and *solcap_snp_c2_25372*, which is located in the citrate synthase encoding locus (PGSC0003DMG400007797) on chromosome 12. Figure 3.3 depicts the physical map of the potato genome sequence (version 4.03) with the 328 significant SolCAP loci from the TSC, TY and TSY case-control studies.
In total, 14 candidate loci from previous reports (Fischer et al., 2013; Li et al., 2008; Urbany et al., 2011, Fischer et al. *in preparation*, Schreiber et al. *in preparation*, Chapter 2) (Appendix Table B.18) were represented in the 328 significant SolCAP SNPs. The relevant loci are shown in Table 3.5.

3.3.3. Association mapping of selected candidate loci

24 polymorphic SNP loci, of which nine were SolCAP SNP markers, were obtained by genotyping six selected candidate loci in the Quest population. The genotypic data are

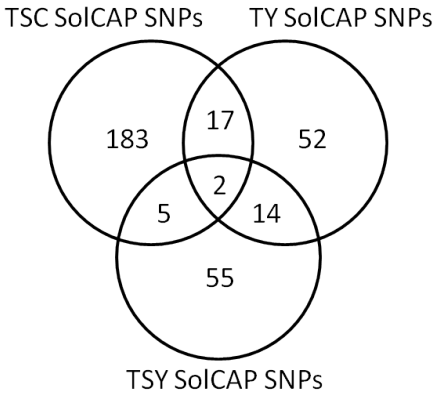


Figure 3.2.: **SolCAP SNPs with significantly (α=0.01) different allele frequencies between pools of case-control studies.** The detailed list of all significant SolCAP SNP loci is provided in Appendix B (Table B.17)

Table 3.5.: **Loci from previous reports (Appendix Table B.18) that were present in the 328 significant SolCAP SNPs**

Locus	Chromo-some	PGSC gene id (PGSC0003...)	Case-control study
G6PPT-1	1	DMG400044320	TSC
Pho1a	3	not on physical map	TSC
AMY-4/2	4	DMG400009891	TSC
SBE I	4	DMG400009981	TSC
CP12-3	4	DMG400009928	TSC
CP12-1	6	DMG400007286	TY
DBE-6/1	6	DMG402007274	TSC
SPS-7	7	DMG400027936	TY
AGPaseB-a	7	DMG400031084	TSC
HSP70	9	DMG400008917	TY
BMY-9	9	DMG400001549	TSC
INV-10/2	10	DMG400008388	TY
INV-n-11/3	11	DMG400026530	TSY
HXK-12	12	DMG400000295	TSY

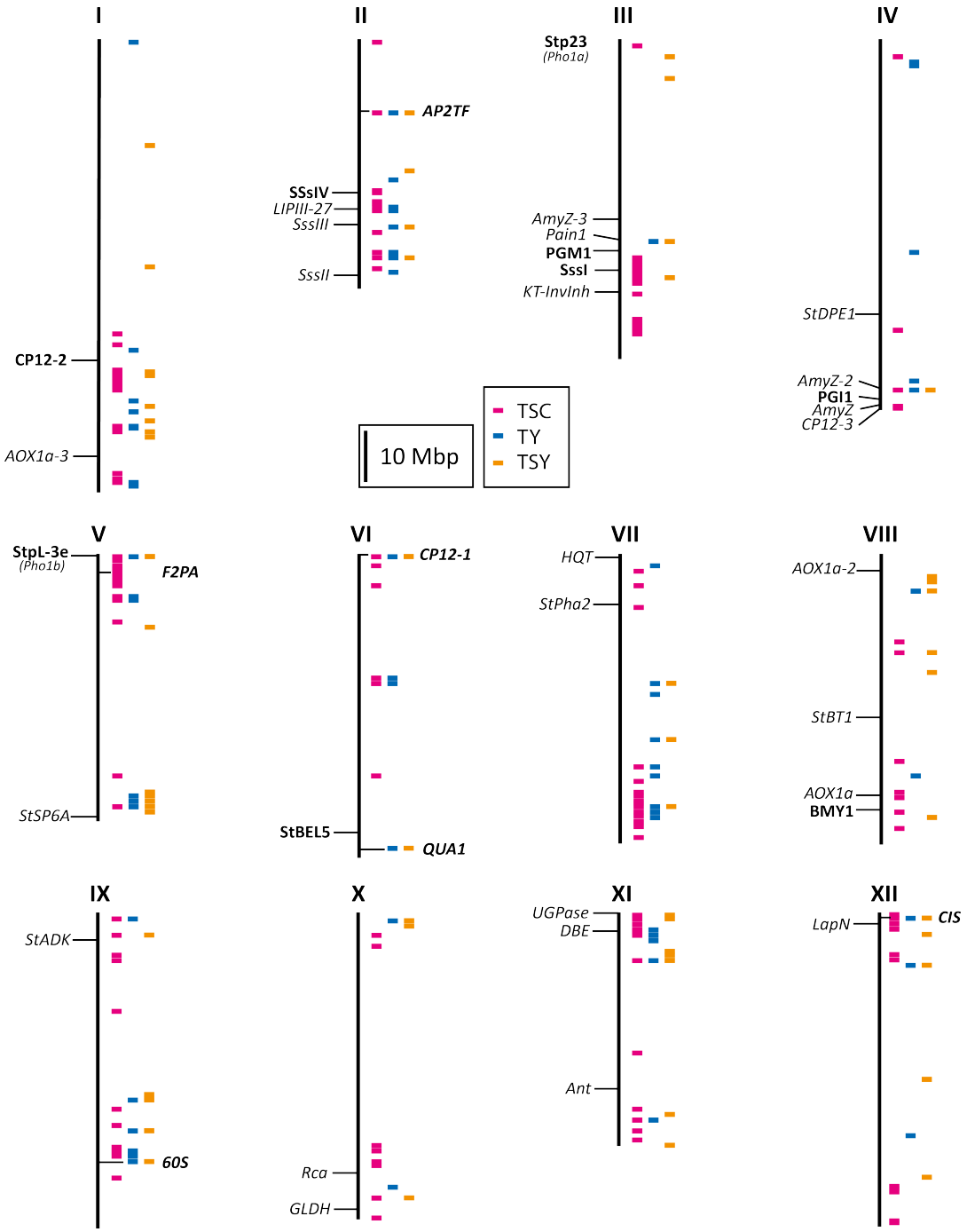


Figure 3.3.: **Physical map of potato (version 4.03).** Shown are significant (α=0.01) SolCAP SNP markers from the case-control studies for TSC, TY and TSY. Candidate loci obtained from the case-control studies with SolCAP Potato Array genotyping that were tested by association mapping in the QUEST population are depicted in bold italics on the right side of the chromosome

given in Appendix B (Table B.7). The polymorphic markers were tested for marker-trait associations with TSC, TY, TSY, TN and TW by implementing a mixed linear model, including population structure and kinship. The association statistics were described in Chapter 2.
Seven polymorphic markers (31.8%) were significantly (α=0.01) associated with at least one phenotypic trait. Two markers were associated with TSC, three markers with TY and three markers with TSY. There were no associations for TN and TW. Marker-trait associations are displayed in Table 3.6. All markers associated with any trait (α=0.05) are reported in Appendix B (Table B.8).

Six SolCAP SNP markers selected from case-control studies The six candidate loci were selected based on significantly different allele frequencies of SolCAP SNps in pools of case-control studies for high and low TSC, TY and TSY. The *AP2TF*, *QUA1* and *CIS* encoding loci were analyzed by amplicon sequencing, where the relevant SolCAP SNP was genotyped along with further SNPs in the amplicon. The SolCAP SNPs of the *F2PA*, *CP12-1* and *60S* loci were analyzed by pyrosequencing. Marker-trait associations with the traits of four out of the six selected SolCAP SNP markers were validated in the full population.
Allele frequencies of the *AP2TF_solcap_snp_c2_16349* marker were significantly different between the pools of all three case-control studies. However, there was no statistical evidence for marker-trait associations in the full QUEST population. *QUA1_solcap_snp_c2_9204* showed significant different allele frequency distribution between pools for high and low TY. Association mapping of the SNP in the full population confirmed this marker-trait association (p-value <0.05). Also the association of *F2PA_solcap_snp_c2_11924* with TSC was verified. Additionally, the polymorphism showed a significant marker-trait association with TSY. Significant marker-trait associations of *60S_solcap_snp_c2_3063* (case-control studies TY, TSY) were detected for TY (α=0.05) and TSY (α=0.01). There was no evidence for an association of *CIS_solcap_snp_c2_25372* (case-control studies TSC, TY, TSY) with any of the phenotypic traits when analyzed in the large panel.

Result of SolCAP SNP markers that were not significant in case-control studies Three SolCAP SNPs that showed no significant difference between allele frequencies in the contrasting pools of the case-control studies were genotyped. This was done either because of the presence in the Sanger sequencing amplicon (*AP2TF_solcap_snp_c2_16350*, *QUA1_solcap_snp_c2_9203*) or the proximity to the SolCAP SNP in the pyrosequencing assay (*F2PA_solcap_snp_c2_11925*). In the case of *QUA1_solcap_snp_c2_9203*, the p-value of the chi-square test in the case-control study for TY was 0.055.

Table 3.6.: **Significant marker-trait associations of candidate loci in the** QUEST **population.** Loci were selected from the case-control studies for TSC, TY and TSY and genotyped by amplicon sequencing or pyrosequencing

Locus	SNP locus	SNP alleles	Minor SNP allele frequency	TSC p-value (R^2)	TY p-value (R^2)	TSY p-value (R^2)	Case-control population
AP2TF	c2_16349[1]	G/A[2]	0.014 (A)[3]	ns[4]	ns	ns	TSC, TY, TSY
F2PA	c2_11924	G/A	0.098 (A)	0.003 (2.9) ↑[5]	0.013 (2.0) –	0.000 (4.3) ↑	TSC
CP12-1	c2_54011	C/T	0.097 (T)	ns	0.004 (2.6) ↓	0.004 (2.8) ↓	TY
QUA1	c2_9203	C/G	0.467 (C)	ns	0.008 (2.2) ↑	0.023 (1.7) ↑	ns
	snp1506	G/A	0.073 (G)	ns	0.004 (2.6) ↑	ns	–
	c2_9204	G/A	0.465 (G)	ns	0.014 (1.9) ↑	0.047 (1.3) ↑	TY
60S	c2_3063	C/T	0.149 (C)	ns	0.015 (1.9) ↓	0.001 (3.6) ↓	TY, TSY
CIS	c2_25372	C/T	0.474 (T)	ns	ns	ns	TSC, TY, TSY
	snp6741	C/G	0.435 (C)	0.001 (3.4) ↑	ns	ns	–

[1]SolCAP SNP locus full name is 'solcap_snp_...' [2]The nucleotide represented in the potato genome sequence stands on the first position; [3]Minor allele frequency of allele displayed in parenthesis; [4]Not significant at significance level α=0.01; [5]Arrows indicate the direction of the minor frequency allele effect on the trait compared to the population mean;

F2PA_solcap_snp_c2_11925 and *QUA1_solcap_snp_c2_9203* showed significant (α=0.01) associations with TY. *QUA1_solcap_snp_c2_9203* was in highly significant LD (q-value=1.54E-205) with the *QUA1_solcap_snp_c2_9204* polymorphism, which was selected from the case-control study for TY. There was no evidence for the association of *AP2TF_solcap_snp_c2_16350* with any trait in the full QUEST population.

Further marker-trait associations Two marker-trait associations were detected between polymorphic SNP loci in the amplicons of *QUA1* and *CIS*, respectively. *QUA1_snp1506* was significantly associated with TY and the presence of the minor frequency allele had a positive effect on the trait. *CIS_snp6741* was significantly associated with TSC with the minor frequency SNP allele having a positive effect on the trait.

3.3.4. Comparing genotyping results

The results of the genotyping with the SolCAP SNP array and the subsequent genotype assignment by 'fitTetra' were compared to the genotyping by amplicon sequencing and pyrosequencing.

The differences between the genotyping by amplicon sequencing and 'fitTetra' ranged between 8-33%. The strongest deviation was detected for *AP2TF_solcap_snp_c2_16349* with 33%. Here, the clustering results of 'fitTetra' showed the genotype classes {0,1,2} and {4}. Manual analysis of the SNP (SNP alleles A/B) showed that signals for *AP2TF_solcap_snp_c2_16349* were seperated in two distinct main genotype classes {ABBB, BBBB}. In line with this result, two genotype classes {ABBB, BBBB} were detected by amplicon sequencing. A comparison of the groups showed that the 'fitTetra' classes {0,1,2} were in one group with amplicon sequenced genotype class {BBBB} and 'fitTetra' class {4} was in one group with amplicon sequenced genotype class {ABBB}. The deviation between the two genotyping methods for the *AP2TF_solcap_snp_c2_16350* polymorphism was 27%. Manual analysis of the SNP showed that the pattern of the SolCAP genotying did not allow genotype assignment and that the genotype assignments of 'fitTetra' for that SNP were artifacts.

The differences between the genotyping by pyrosequencing and the 'fitTetra' output ranged between 0-6%, except for *F2PA_solcap_snp_c2_11925* (55%). The clustering results of the *F2PA_solcap_snp_c2_11925* SNP by 'fitTetra' showed the genotype classes {0}, {2} and {4}, whereas all five genotype classes were found with pyrosequencing. By manual clustering it was very hard to separate the heterozygotes from each other. As a result of that, 'fitTetra' clustered all heterozygotes in class {2}, corresponding to genotype {AABB}.

Table 3.7.: **Comparing genotyping by the SolCAP Potato Array based on genotype assignment of 'fitTetra' with Sanger amplicon sequencing and pyrosequencing**

Locus	SolCAP SNP (solcap_snp_...)	Total	Number of genotypes		Different in [%]
			identical	different	
Amplicon sequencing					
AP2TF	*c2_16349*	90	60	30	33
	c2_16350	90	66	24	27
QUA1	*c2_9203*	89	82	7	8
	c2_9204	89	70	19	21
CIS	*c2_25372*	89	77	12	13
Pyrosequencing					
F2PA	*c2_11924*	85	84	1	1
	c2_11925	87	39	48	55
CP12-1	*c2_54011*	90	90	0	0
60S	*c2_3063*	84	79	5	6

3.3.5. Linkage disequilibrium

LD between significantly different SolCAP SNPs from case-control studies A chi-square test was performed to test the LD between the alleles of the 328 SolCAP SNP markers that showed significantly different allele distributions between contrasting pools in at least one case-control study (Appendix Table B.12). Of the total of 53,628 pairwise comparisons, 5.7% were in significant LD (q-value <0.05). Of all intra-chromosomal marker pairs 14.5% were in significant LD (q-value <0.05), whereas the amount of inter-chromosomal marker pairs in significant LD in relation to the total amount of inter-chromosomal marker pairs was 4.8%.

LD clusters within chromosomes 1 and 3 The highest LD was found in intra-chromosomal regions on chromosomes 1 and 3, respectively, where large clusters of marker pairs in LD were detected.

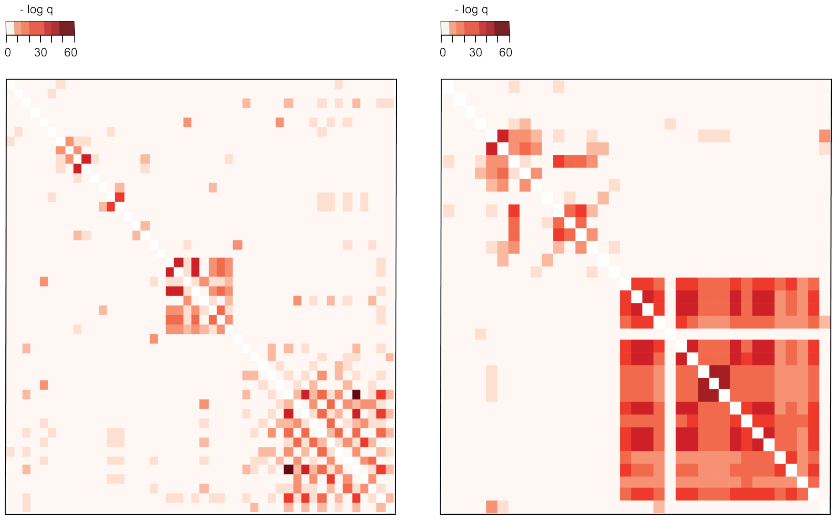
The largest LD cluster was detected on chromosome 3. LD between pairs of markers for chromosome 3 is shown in Figure 3.4 (b). The obvious cluster of SolCAP SNP markers in strong LD was located in the region between 54.1-56.7 Mbp (chr03:54141884..56672205). The physical distance between the flanking markers (*solcap_snp_c2_47843*, *solcap_snp_c1_8194*) was 2.53 Mbp. There were neither QTL markers nor candidate genes from previous studies located in that region of the chromosome. The *Pain1* locus was also not part of the the LD cluster on chromosome 3. Furthermore, the *PGM1* and *SssI* loci were not located in the LD cluster. All marker pairs in highly

significant (q-value <1,00E-06) LD were obtained from the case-control study for TSC. Schreiber et al. (*in preparation*) presented a detailed list of 127 genes that were mapped *in silico*. All genes are operational in starch synthesis and degradation and where cloned and characterized in either potato or tomato. The genome sequence of the region contained many genes as well as two large assembly gaps. However, the LD cluster did not contain any of the listed genes.
Furthermore, two smaller clusters of marker pairs in high LD were identified on chromosome 1. The larger cluster of the two was located at the distal end of chromosome 1, flanked by SolCAP SNPs *solcap_snp_c2_14730* and *solcap_snp_c2_30956* (chr01:86441447..88451942, length 2.01 Mbp). This region was located in 0.4 Mbp distance from the *AGPaseS-a* locus, which is located in a QTL for TSC (Schäfer-Pregl et al., 1998). Looking at all marker pairs that are in highly significant (q-value <1,00E-06) LD in the cluster, significant markers from all three case-control studies for TSC, TY and TSY were represented. Of the gene list from Schreiber et al. (*in preparation*), the *INV-1/3* (PGSC0003DMG400001596) was located in that region, containing three SolCAP SNP markers. No different allele frequencies were detected for these three polymorphisms in any of the case-control studies. The second marker cluster on chromosome 1 was located in the region around 75 Mbp (chr01:75333643..76958587), stretching over a length of 1.6 Mbp between *solcap_snp_c1_5267* and *solcap_snp_c2_2423*. Figure 3.4 (a) depicts the LD plot of markers on chromosome 1. The region included the *GP88* marker for a yield QTL (Schäfer-Pregl et al., 1998). Significant markers from all three case-control studies for TSC, TY and TSY were present in this region.

LD between SNPs of selected candidate loci LD between all pairs of the 23 genotyped SNPs from candidate loci was tested with a chi-square test. In total, 67 (26.5%) out of 253 marker pairs were in significant (q-value <0.05) LD. The highest LD was found between markers from the same locus (Figure 3.4 c, Appendix Table B.11).

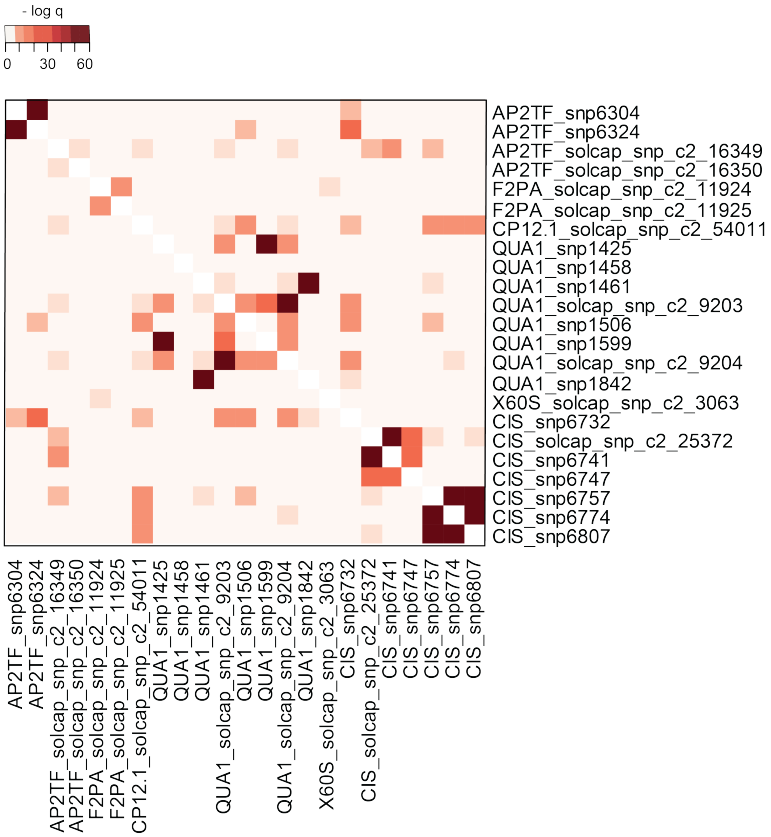
3.3.6. Comparing mixed model and general linear model

The most significant marker-trait associations of three statistical models are summarized in Table 3.8: the results of the general linear model (GLM), the mixed linear model including kinship (MLM-k) and the most stringent mixed linear model including kinship and population structure (MLM-Pk). The comparison between the models is given in Appendix B (Table B.9).



(a) Chromosome 1

(b) Chromosome 3



(c) SNP alleles of candidate loci

Figure 3.4.: **Linkage disequilibrium between marker pairs of significant SolCAP SNP loci on chromosomes 1 (a) and 3 (b) as well as between SNP alleles of candidate loci that were scored in the entire population (c).** Displayed are -log q-values of chi-square test between pairs of SolCAP SNPs. Each row and line corresponds to one SNP position

Table 3.8.: **Markers that were highly significantly associated in three different models.** Significance levels were α=0.001 in general linear model (GLM) and mixed linear model including kinship (MLM-k) and α=0.01 for mixed linear model including kinship and population structure (MLM-Pk). Markers with p-values <0.0001 in GLM and MLM-k are considered "robust" and are shown in bold

SNP locus	Model	TSC	TY	TSY
CIS_snp6741	GLM	**5.98E-06**		
	MLM-k	**9.67E-06**		
	MLM-Pk	**0.001**		
CP12-1_solcap_snp_c2_54011	GLM		**4.84E-08**	**4.14E-05**
	MLM-k		**2.05E-07**	**8.81E-06**
	MLM-Pk		**0.004**	**0.005**
F2PA_solcap_snp_c2_11924	GLM	4.08E-04		
	MLM-k	5.10E-04		
	MLM-Pk	0.003		
QUA1_solcap_snp_c2_9203	GLM		7.48E-05	
	MLM-k		1.13E-04	
	MLM-Pk		0.008	
QUA1_snp1506	GLM		1.26E-04	
	MLM-k		1.82E-04	
	MLM-Pk		0.004	
60S_solcap_snp_c2_3063	GLM			**1.04E-05**
	MLM-k			**1.96E-05**
	MLM-Pk			**0.001**

3.4. Discussion

3.4.1. Novel diagnostic SNP markers detected by genotyping in a case-control design

Case-control studies are a suitable tool for the pre-selection of non-obvious candidate genes for developing diagnostic SNP markers. Ninety tetraploid potato genotypes were selected from an association mapping population that was phenotyped for starch and yield-related traits. In total, three case-control studies were designed for TSC, TY and TSY. All individuals were genotyped with 8,303 SNP markers by the SolCAP Potato Array and genotypes were assigned using the 'fitTetra' package (Voorrips et al., 2011). Statistical analysis of the SNP allele frequency between contrasting pools of case-control studies revealed 328 SNPs with significantly different allele frequencies between pools. Six SolCAP SNPs that had significantly different allele frequencies in contrasting pools for TSC, TY or TSY, or a combination of those, were genotyped in the full association mapping population. Association mapping could validate marker-trait associations of four out of the six selected SolCAP SNPs. The markers showed highly significant associations with the traits which is probably based on the strong pre-selection of the candidate SolCAP SNP loci. By that, it was possible to test markers in loci that are non-obvious candidate genes for the complex traits. It is unlikely that the loci would have been chosen in a knowledge-based approach. In the light of the extend of LD reported for tetraploid potato cultivars and breeding material (e.g. D'hoop et al., 2010; Simko et al., 2006; Stich et al., 2013), it remains unclear whether the identified markers are located directly in a causal gene or more likely in LD with the causal gene. The case-control study design, which has been adopted from human disease studies, has proven to be a valuable tool for the dissection of complex traits in tetraploid potatoes. It is a straightforward method for the pre-selection of candidate genes that should be further confirmed in a larger genetic background.

Markers for tuber starch yield and tuber yield The complex traits tuber yield and starch yield may be influenced by enzymes that are active in pathways that regulate sugar levels in plants. An interesting candidate is the Calvin cycle chloroplast protein 12 locus (*CP12-1*) on chromosome 6. In the case-control study for TY, *CP12-1_solcap_snp_c2_54011* showed significantly different allele frequencies between the contrasting pools. CP12 was proposed to have a functional role in potato yield, shown in an RNA sequencing experiment (Kare Lehman Nielson, EAPR meeting 2010, oral presentation). Furthermore, an association of the *CP12-2* locus with TN, a yield-related trait, was found in the QUEST population (Chapter 2). In this study, *CP12-1_solcap_snp_c2_54011* was highly significantly associated

with TY and TSY in all three statistical models in the QUEST population. CP12 was identified to interact with the Calvin cycle proteins GAPDH (glyceraldehyde 3-phosphate dehydrogenase) and PRK (phosphoribulokinase) (Pohlmeyer et al., 1996; Wedel et al., 1997). The Calvin cycle is responsible for the carbon fixation in plants and located in the chloroplasts. Moreover, the *CP12* genes are expressed in non-photosynthetic tissues in *A. thaliana* (Singh et al., 2008), which is in line with the constitutive expression of *CP12-1* in potato tissues (genome browser version v.4.03). These findings suggest a role of *CP12-1* in non-photosynthetic tissues in potato. The expression of *CP12-1*, presumably in amyloplasts, may regulate GAPDH that is also active in glycolysis. Sequence variation may increase the activity or expression of the protein, enhancing the availability of plant sugars and thereby indirectly influencing yield levels. This makes *CP12-1_solcap_snp_c2_54011* a highly beneficial marker for the development of superior cultivars with optimized tuber- and starch yield.

A SolCAP SNP at the non-obvious candidate locus encoding a predicted 60S acidic ribosomal protein PO (*60S*) is of potential value for marker-assisted selection for starch yield in breeding programs. The *60S_solcap_snp_c2_3063* showed significantly different allele frequencies in the TY and TSY case-control studies. When genotyped and tested in the QUEST population, these marker-trait associations were validated and highly significant in all three statistical models. The minor frequency allele had a negative effect on the trait with a frequency of about 15% in the population. In a breeding program the allele frequency should be reduced to obtain cultivars with potentially higher yield and starch yield levels. *60S* is located in a region with no described QTL for starch or yield. Still, the candidate SNP was selected based on its putative function in a ribosomal subunit. Polymorphisms could potentially influence functionality and thereby fitness of the plant. *60S_solcap_snp_c2_3063* SNP could therefore be of high value for the selection of potato cultivars with improved tuber and starch yield.

Markers for tuber starch content The association of the *F2PA_solcap_snp_c2_11924* polymorphism in the fructose-bisphosphate aldolase encoding *F2PA* locus on chromosome 5 with TSC and TSY may be a result of increased source capacity. The SNP was selected from the output of the case-control study for TSC. It co-localizes with a starch and yield QTL on the top of chromosome 5 (Schäfer-Pregl et al., 1998). The *F2PA* locus has sequence homology with the recently cloned *Solanum tuberosum* plastidial aldolase (NM_001288043). Plastidial F2PA is an enzyme of the Calvin cycle and has an effect on photosynthetic carbon fixation (Uematsu et al., 2012). Jonik et al. (2012) showed that transgenic potato plants with increased source-capacity in combination with elevated sink strength resulted in elevated levels of starch content and yield. Therefore, the associations

of *F2PA_solcap_snp_c2_11924* with TSC and TSY might by explained by the activity of F2PA in the Calvin cycle. Sequence variation in *F2PA* may lead to a different expression of haplotypes or a change in protein composition, thereby enhancing the carbon fixation rate and indirectly elevating starch content by increased source capacity.
A valuable marker for starch yield optimization in breeding programs is a novel polymorphism in the citrate synthase *CIS* locus on chromosome 12. One SNP of the candidate locus (*CIS_solcap_snp_c2_25372*) was associated in all three case-control studies for TSC, TY and TSY. Although these associations were not validated by association mapping, a marker-trait association at the *CIS* locus (*CIS_snp6741*) was found highly significant with TSC. Citrate synthase is the first enzyme of the citric acid cycle. In the citric acid cycle, the plant generates energy as well as precursors for numerous biochemical reactions (Wiegand and Remington, 1986). The energy production is maintained by the degradation of carbohydrates and other products (Wiegand and Remington, 1986). An explanation for the association at the *CIS* locus with TSC may be a reduced activity of *CIS* alleles, resulting from sequence variation, which may indirectly maintain levels of carbohydrates in sink tissues and thereby starch content. Based on the highly significant marker-trait association in all statistical models and the function of the gene, *CIS_snp6741* is a beneficial SNP marker for the development of cultivars with elevated levels of TSC.

Markers for tuber yield *QUA1* encodes a glycosyltransferase QUASIMODO1 protein. Two polymorphisms in this non-obvious candidate gene are potential markers for TY. *QUA1_solcap_snp_c2_9204* had significant different allele frequencies between the pools of the TY case-control study. The *QUA1_solcap_snp_c2_9203* polymorphism lies in the same locus and was not significant in the case-control study for TY. However, it is in almost complete linkage with *QUA1_solcap_snp_c2_9204* and it was close to significance in the case-control study for TY (p-value=0.055). Both polymorphisms were significantly associated with TY in association mapping. Opposed to *QUA1_solcap_snp_c2_9204*, *QUA1_solcap_snp_c2_9203* is highly significant in all three statistical models, with a higher amount of variance explained by the model. This makes *QUA1_solcap_snp_c2_9203* a good candidate for marker-assisted selection for superior genotypes with high TY levels. The minor frequency allele has a positive effect on the trait and its frequency should be increased in a breeding program. Similarly, increasing the minor frequency allele of *QUA1_snp1506* by positive selection can potentially enhance yield levels. The polymorphism was detected by amplicon sequencing of the *QUA1* locus and shows a significant and positive marker-trait association with TY. Two valuable SNP markers for TY (*QUA1_solcap_snp_c2_9203* and *QUA1_snp1506*) were obtained from the non-obvious candidate *QUA1*.

Not all associations from case-control studies could be verified in the entire population
However, not all SolCAP SNPs from case-control studies lead to significant marker-trait associations when tested in the full QUEST population. *CIS_solcap_snp_c2_25372* was associated in all three case-control studies, while there was no evidence for marker-trait associations in association mapping. This result is in line with the report of Huang et al. (2011), who found that loci identified by genome-wide genotyping in smaller population sizes are not necessarily significant in a larger populations. Similarly, *AP2TF_solcap_snp_c2_16349* was significant in all three case-control studies without evidence for marker-trait associations in the full population. However, there was a large deviation between the genotyping of *AP2TF_solcap_snp_c2_16349* by 'fitTetra' on the one hand and by amplicon sequencing on the other hand. This proved to be a result of incorrect clustering by 'fitTetra', leading to inflated allele frequencies and a false-positive in the analysis. When comparing the results of the case-control studies with association mapping in the full size population, missing associations can be caused by the fact that not all loci that are associated in small panels are also associated in larger populations, but might also be due to genotyping errors.

3.4.2. Genotyping with the SolCAP Potato Array appropriate tool for marker detection in populations of European potato germplasm

The SolCAP Potato Array is a suitable tool for the large scale genotyping of European potato cultivars and breeding clones. 6,065 polymorphic SNP markers of totally 8,303 tested markers were informative. The success rate of the SolCAP genotyping (73%) is comparable to the results of a previous experiment by Stich et al. (2013). They reported a success rate of 75% after genotyping 36 tetraploid european potato cultivars with subsequent manual assignment of genotypes. Both studies confirm that the application of the SolCAP Potato Array is highly beneficial for the genotyping of commercial European potato germplasm.
However, the genotyping with the SolCAP Potato Array has a bias towards polymorphisms of North American potato germplasm. A limitation of the SolCAP Potato Array is that the polymorphisms on the array were primarily obtained from North American cultivars (Hamilton et al., 2011). Therefore, polymorphisms that separate European potato genotypes might not be present on the SolCAP Potato Array. Furthermore, the SNPs were discovered between only five genotypes and polymorphisms tagging minor allele frequency were probably not detected. Uitdewilligen et al. (2013) showed that the detection variant with such a small number of genotypes is likely limited to the most common variants. An indication for the under-representation of minor frequency alleles and polymorphisms that are characteristic for European germplasm is the unexpectedly low number of significant candidate loci from starch and yield-related traits from previous studies, despite the pres-

ence of SolCAP SNPs in these loci. Thus, when genotyping potato clones with the SolCAP Potato Array one has to keep in mind that there is a potential under-representation of genetic variation that is characteristic for European potato genotypes. If the focus would lie on the latter, the recently developed 20k array (SolSTW-20k-Wageningen Uitdewilligen, 2012) might be a suitable choice for genotyping.

3.4.3. Clusters of SolCAP SNP loci in highly significant LD are likely result of limited meiotic recombination between potato lines

The large haplotype blocks could be a consequence of the limited meiotic recombinations between potato genotypes. A region on chromosome 3 was detected with significant SNPs from the case-controls study for TSC that were in very high LD. The markers are spread over 2.53 Mbp. This is a clear deviation from the reported LD decay between 275 bp (Stich et al., 2013) and 70 kbp (Simko et al., 2006). As all significant SolCAP SNPs in that region were solely obtained from the TSC case-control study, one explanation is that this region contains genes and haplotypes of importance for TSC. The Quest population was assembled with the focus on starch properties, possibly enhancing the amount of genotypes containing the LD cluster. Another explanation is that population structure might have had an influence on the the result of the case-control studies, as suggested by Pritchard and Donnelly (2001). One can imagine, that the individuals within pools are likely to be more closely related than individuals between pools. Although a modest population structure was detected in the Quest population, there is no indication to assume that an underlying structure could have influenced the result. The reason is, that the substructure was separating tetraploid cultivars and breeding clones from landraces and landraces were not selected for the case-control studies. Nevertheless, no starch QTL have been reported for that region up to now and no obvious candidate genes from the starch synthesis or degradation are located in that region. The existence of large haplotype blocks were also reported in previous association mapping populations that were assembled from potato varieties and advanced breeding lines (Gebhardt et al., 2004; Malosetti et al., 2007; Simko et al., 2006) and might be a result of selection or the limited amount of meiotic recombination between lines. Similarly, two LD clusters were detected on chromosome 1, with significant SolCAP SNP markers from all three case-control studies. There is a possibility that the LD blocks are the result of selection. Due to the lack of any obvious candidate genes or QTL from previous studies, the LD blocks are more likely to be a consequence of the low genetic distance between cultivars.

3.4.4. Usefulness and limitations of case-control studies for dissecting quantitative traits in potatoes

The advantages of case-control studies for discovering candidate loci associated with quantitative traits in potato are manifold, although the method has its limitations. Applying a case-control study design for the dissection of quantitative traits has three major advantages. Firstly, the populations are easy to establish. Pools of genotypes with available phenotype data for the desired trait can be assembled directly. The method does not require the generation of a bi-parental mapping population, like in the case of bulked segregant analysis. Secondly, by the strong bias towards the trait of interest, markers for implementation in breeding programs can be developed quickly. A genome-wide genotyping approach allows the detection of novel and non-obvious candidate genes. However, case-control studies also have their limitations. It remains difficult in the application of a pooling strategy to distinguish if the candidate genes are causal genes, in LD with the causal gene or simply false-positives from the analysis (Kloosterman et al., 2010). Furthermore, hidden population structure might influence the result of case-control studies, especially if very diverse germplasm is used (Pritchard and Donnelly, 2001). Keeping in mind the limitations, the case-control study design is highly beneficial for the dissection of complex traits in potato.

3.4.5. Concluding paragraph

In this study, a case-control study design in combination with genome-wide SNP genotyping was tested for applicability in development of novel diagnostic SNP markers for complex traits. The concept of the case-control study design was adopted from human genetics. In alliance with the SolCAP Potato Array genotyping, it is a comparably low-cost genome-wide genotyping concept that is time saving and results in novel, non-obvious candidate genes. The results prove that the described approach is of great interest for the detection of diagnostic SNP markers.

4. Genome-wide SNP discovery by RAD sequencing in tetraploid potato unravels novel candidate genes for starch yield traits

4.1. Background

The dissection of the genetic factors underlying the variation in phenotypic traits is of major interest in genetic studies. A prominent example in potato is Sanger amplicon sequencing of candidate genes and subsequent association mapping (Chapter 2). A limitation of that method is that the selection of appropriate candidate genes is based on knowledge that is present in literature. Furthermore, only a limited number of sequence variants can be assessed with this method as primer design is challenging for the highly heterozygous tetraploid potato. Increased numbers of markers are obtained by higher-density genotyping, which became possible with the availability of the SolCAP Potato Array (Chapter 3). Still, probes on this array were designed by knowledge-based selection criteria (Felcher et al., 2012).
The unbiased genome-wide detection of markers became feasible with the development of next-generation sequencing methods that reduce the complexity of genomes and the amount that needs sequencing. RAD sequencing is a tool for genome-wide genotyping of DNA polymorphisms. The polymorphisms are randomly distributed over the whole genome and therefore novel, non-obvious candidate genes can be detected. Urbany et al. (2012), Fischer et al. (2013) and the study described in Chapter 3 of this thesis showed that a case-control study design is a valuable tool for marker-development for complex traits in potato.

Approach The main objectives were to test the applicability of RAD sequencing for large-scale genotyping of tetraploid potato and to identify novel candidate genes for starch and tuber yield traits. In order to maximize the output at a limited amount of costs, the case-control design was applied in the combination with RAD sequencing.

4.2. Materials and Methods

4.2.1. Experimental design

Three case-control studies with phenotypically contrasting pools for tuber starch content (TSC), tuber yield (TY) and tuber starch yield (TSY) were established from tetraploid varieties and breeding clones of the QUEST population. The background of the population and the selection of phenotypes for the case-control studies was described in detail in Chapters 2 and 3, respectively. A total of ninety genotypes were represented in the three case-control populations. Additionally, six technical replicates were included in the experiment.

Assumptions on restriction sites, RADtags and SNP detection The potato genome has a size of 844 Mb with a GC content of 35% (The Potato Genome Sequencing Consortium, 2011). The recognition site of the selected restriction enzyme *Kpn*I is G'GTACC. Based on both genome size and GC content, the estimated number of restriction enzyme recognition sites (restriction sites) in the genome was 82,219. This corresponds to a total amount of 164,439 RADtag, as two RADtags are created at each restriction site (Table 4.1).
Aiming at a an average fold coverage of 5x of each of the four alleles per genotype, 48 genotypes could be multiplexed per sequencing lane for paired-end sequencing on an Illumina HiSeq2000. On average, 300 million sequencing reads are generated per sequencing lane. The generated sequence length per RADtag was expected to be about 80 bp after removing adapter and barcode sequence. Based on a size selection of about 300-400 bp (see library preparation), the paired-end sequence was generated in 300-400 bp distance of the recognition site. Therefore, the sequence generated per RADtag was expected to be about 160 bp. The reported SNP frequencies of potato differed between 1 SNP/140 bp in exon regions, 1 SNP/80 bp in intron regions (Uitdewilligen, 2012) up to 1 SNP/21 bp Rickert et al. (2003), which is similar to the findings of Uitdewilligen et al. (2013) (1 SNP/24 bp in exon and 1 SNP/15 bp in intron regions). This lead to assumed numbers of 183,359, 320,878 and 1.22 million detected SNPs, respectively (Table 4.1).

4.2.2. RAD library preparation and sequencing

The RAD libraries for paired-end sequencing were prepared according to Etter et al. (2011) with modifications by Bus et al. (2012). In brief, 2 ng of genomic DNA of each genotype was digested with the *Kpn*I restriction enzyme. The 96 samples were individually barcoded with a custom-made P1 adapter (Bus et al. *unpublished data*), which was ligated to the restriction enzyme cut sites. The samples were then pooled into two libraries of 48 individuals each. Samples were sheared by ultrasound and fragments of 300-400 bp

Table 4.1.: **Assumptions on restriction sites, RADtags and SNP detection.**

Detail	Number
Size of potato genome[1]	844 Mb
GC content potato genome	35%
Expected number of *Kpn*I restriction sites	82,219
Expected number of RADtags	164,439
Illumina HiSeq2000 sequences per lane	300,000,000
Expected sequence per RADtag minus adapter	160 bp
SNP frequencies and expected number of SNPs	
1 SNP/140 bp exon[2]	183,359
1 SNP/80 bp intron[2]	320,878
1 SNP/21 bp[3]	1,252,869

[1] The Potato Genome Sequencing Consortium (2011); [2] Uitdewilligen (2012); [3] Rickert et al. (2003)

length were selected by extraction from an agarose gel. An A-overhang was added to the blunt ends and the common P2 adapter (Bus et al. *unpublished data*) was ligated to the template. After a PCR enrichment step, a final size selection was performed. The detailed protocol for the RAD sequencing library preparation in potato is given in Appendix B (Protocol B.2). Library quality was determined on a 2100 Bioanalyzer (Agilent Technologies, Böblingen, Germany) and the two libraries were sequenced in two lanes of the Illumina HiSeq2000 system at the Max Planck-Genome-centre Cologne (Germany) with GAIIx chemistry.

The adapter ligation steps were optimized and P1 as well as P2 Adapters were ligated at 16°C overnight instead of 30 min at room temperature, due to a re-occurring contamination band at 150 bp. The 96 barcoded P1 adapters, the common P2 adapter as well as PCR primers for the enrichment step were provided by Anja Bus (Bus et al. (*unpublished data*). The barcodes of the P1 adapters were designed to be at least six mutational steps separated from each other.

4.2.3. Sequence analysis and SNP detection

The sequence analysis and SNP detection as well as statistical analyses were performed by Jia Ding (MPIPZ, Cologne, Germany). Paired-end reads were combined from both sequencing lanes and reads were sorted according to their barcode, allowing the maximum of 1 mismatch. Reads with a barcode showing more than one mismatch were discarded. No mismatch was allowed at the restriction site. The sequences were mapped against the potato genome sequence (version v2.1.11) (The Potato Genome Sequencing Consortium, 2011) using the Bowtie package (Bowtie; Langmead et al., 2009). Reads that did not map

to a unique position in the genome were excluded from further analyses.
Sequencing reads of individual genotypes were assigned to the three case-control populations for TSC, TY and TSY, accordingly. Further analysis was performed for pooled sequence reads. Bi-allelic SNP positions were called within pools using the Genome Analysis Toolkit (GATK; McKenna et al., 2010).

4.2.4. Statistical analyses of case-control studies

Fisher's exact test was implemented with a custom-made Perl script to test if allele frequencies of the two SNP alleles were significantly different between the 'high' and 'low' pool of each of the three case-control studies. The result was corrected for multiple testing by FDR (Bonferroni's approach) and significant (FDR <0.05) SNPs were combined with annotation information (version v2.1.11) to obtain a final list of SNP markers in annotated loci. Synonymous and non-synonymous SNPs were identified with the SnpEff package (SnpEff; Cingolani et al., 2012), by comparing the detected SNPs to the annotation file of the potato genome sequence (version v2.1.11).

4.3. Results

4.3.1. Sequencing result and mapping

Sequencing the RAD libraries with totally 96 individually barcoded genotypes on two lanes of an Illumina HiSeq2000 system generated 346 million paired-end reads (Table 4.2). The sequences were processed and statistically analyzed by Jia Ding (MPIPZ, Cologne, Germany). 89% contained an individual 12 bp barcode with the maximum of one mismatch allowed. 40% of the reads had a restriction site, with no mismatch. In total, 139 million reads had an individual barcode and a perfect restriction site. These paired-end reads were mapped to the potato genome sequence (version v2.1.11). Reads that did not map to unique positions in the genome were excluded and the uniquely mapped paired-end sequences were assigned to the case-control studies, accordingly. Further analysis was performed for pooled sequence reads.

Table 4.2.: **Summary statistics for RAD sequencing of 96 genotypes of the** Quest **population.** Sequences were generated on an Illumina HiSeq2000

Statistic	Number (x10^6)
Total Illumina paired-end sequences	346.52
Reads containing barcode ($\leq$1 mismatch)	310.76
Reads containing restriction site	140.33
Reads containing barcode ($\leq$1 mismatch) + restriction site	139.00

4.3.2. Statistical analysis of case-control studies

The total number of SNPs detected in the case-control study for TSC was 579,352. 601,837 SNPs were detected in the case-control study for TY and 598,703 for TSY (Table 4.3.
The SNPs of each case-control study were tested for differences in SNP allele frequencies between the contrasting pools with a Fischer's exact test. p-values were corrected for multiple testing (FDR). In total, 58,850 (TSC), 46,542 (TY) and 20,402 (TSY) SNPs with significantly (FDR <0.05) different allele frequencies between pools were detected. 26% (TSC), 25% (TY) and 29% (TSY) of the significant SNPs were located in annotated loci of the potato genome sequence (version v2.1.11) (Table 4.3). The significant (FDR <0.05) SNPs in the case-control study for TSC were distributed over 5,281 annotated loci. In the case-control studies for TY and TSY, the significant SNPs were distributed over 4,563 and 2,982 loci, respectively.
The effects of the significant SNPs on the protein were predicted with the SnpEff package (Cingolani et al., 2012). 2,793 SNPs (in 1,621 loci) from the case-control study for TSC

had a non-synonymous effect. From the significant SNPs detected in the case-control study for TY the number of SNPs with a non-synonymous effect was 2,139 (in 1,319 loci) and in the case-control study for TSY 1,094 (in 752 loci) (Table 4.3.

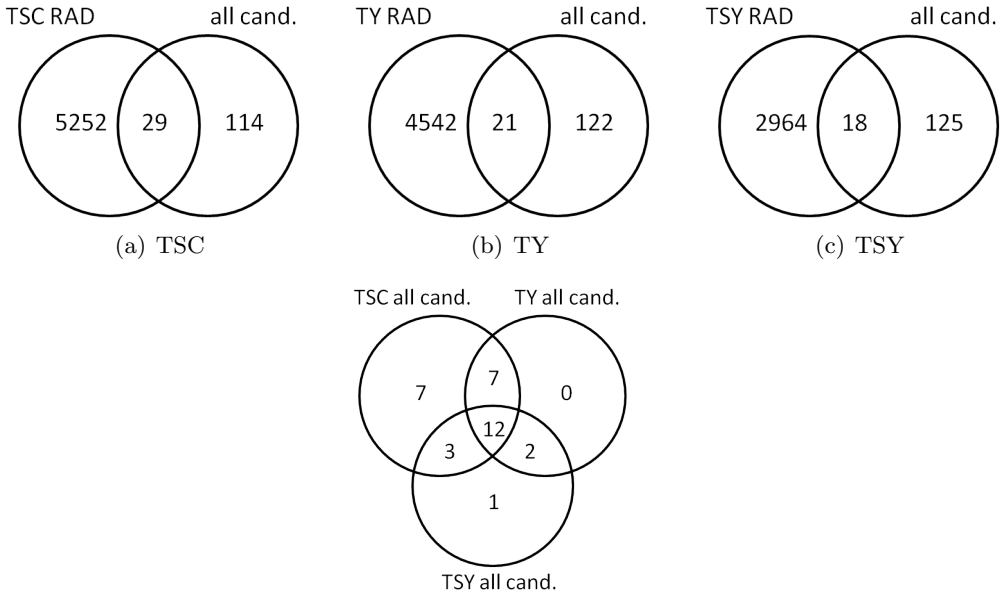
Table 4.3.: **Results of statistical analysis of the case-control studies.** (i) Total numbers of detected SNPs in pools, (ii) SNPs with significantly (FDR <0.05) different allele frequencies between pools as well as (iii) non-synonymous SNPs between pools. The genotyping was performed by RAD sequencing. TSC=tuber starch content, TY=tuber yield and TSY=tuber starch yield

Study	Total	Significant SNPs			Loci with significant SNPs	
		total	in annotated loci		total	non-synonymous
			total	non-synonymous		
TSC	579,352	58,850	15,307	2,793	5,281	1,621
TY	601,837	46,542	11,646	2,139	4,563	1,319
TSY	598,703	20,402	5,850	1,094	2,982	752

4.3.3. Knowledge-based validation of case-control study results

The presence of candidate genes and markers that were described in previous studies (Fischer et al., 2013; Li et al., 2008, 2013; Urbany et al., 2011, Fischer et al. *in preparation*; Schreiber et al. *in preparation*; Chapter 2) (Appendix Table B.18) was investigated in loci that contained at least one significant (FDR <0.05) SNP in the case-control populations. The results of the comparisons are shown in Figure 4.1 (a-c). 29 loci were in common between the case-control study for TSC and the compiled list of candidate genes and markers, 21 loci between the case-control study for TY and 18 loci for TSY. In total, 32 loci detected in any of the case-control studies were represented on the compiled list of 142 candidate genes and markers (Figure 4.1 d, Table 4.4).

PCR markers from previous studies In total, two loci containing allele specific PCR markers from previous studies were detected in the loci of the case-control studies that contained at least one significantly (FDR <0.05) different SNP. The *Rca* locus (PGSC0003DMG400019149) had 4 different SNPs between contrasting pools in the case-control study for TSC. The *Rca-1a* marker was reported to be associated with chips quality (Li et al., 2008) and associated with TSC (Chapter 2). The 4 SNPs detected by RAD sequencing were not included in the *Rca-1a* PCR fragment.
The *HSP70* locus (PGSC0003DMG400008917) had 3 significantly different SNPs in the case-control study for TSC, 7 significant SNPs in the case-control study for TY and 1 significant SNP in the case-control study for TSY. The *HSP70-bad* allele was reported to



(a) TSC (b) TY (c) TSY

(d) intersecting set

Figure 4.1.: **Venn diagrams of the loci detected by RAD sequencing (at least one significant SNP per locus) and that were present on the compiled list of candidate genes and markers from previous studies for the case-control studies (Appendix Table B.18).** (a) TSC, (b) TY, (c) TSY and (d) all overlapping loci detected in any of the case-control studies

Table 4.4.: **Loci from previous reports (Appendix Table B.18) that were represented by at least one significant (FDR <0.05) RAD sequencing SNP and present in at least two case-control populations (Figure 4.1).** Loci with more than five significant RAD sequencing SNPs, containing minimum one SNP with non-synonymous effect, and the relevant case-control study are shown in bold

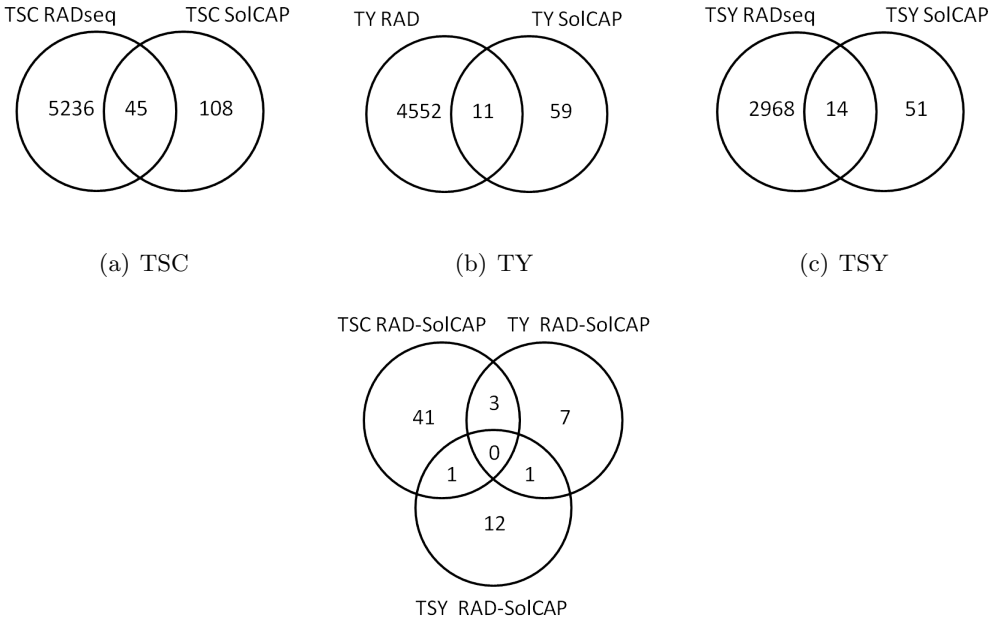
Locus	Chromosome	PGSC gene id (PGSC0003...)	Case-control study
Cis	1	DMG400028982	TSC, TY
AOX1a-3	1	DMG400012558	TSC
INV-1/2	1	DMG400001596	TSC, TY
LIPIII-27	2	DMG400031758	TY, TSY
HT-2/1	2	DMG400022402	TSC, TY, TSY
SuSy-2	2	DMG400016730	TSC
HT-2/2	2	DMG400026402	TSC, TY, TSY
SS V	2	DMG400030619	TSC
HXK-2	2	DMG400030624	TSC, TY, TSY
SssI	3	DMG402018552	TSC, TSY
Pk	4	DMG400025298	TSC, TY
MT	4	DMG400024812	TSC
SBE I	4	DMG400009981	TSC, TY, TSY
GWD	**5**	**DMG400007677**	**TSC**, TSY
239E4left	5	DMG400015743	TSC, TY, TSY
DBE-6/1	**6**	**DMG402007274**	**TSC**, **TY**, TSY
SPS-7	7	DMG400027936	TSC, TY
StPha2	7	DMG400004101	TSC, TY
SuSy-7/2	7	DMG400013546	TSC, TY, TSY
AGPaseB-7	7	DMG400031084	TSC
AOX1a-2	8	DMG400018484	TSY
BMY-8/2	8	DMG400001855	TSC, TY
INV-8/2	8	DMG400004790	TSC, TY, TSY
BMY-8/3	8	DMG400012129	TSC, TY
HSP70	**9**	**DMG400008917**	TSC, **TY**, TSY
StADK1	9	DMG400027906	TSC
HT-9	**9**	**DMG400031832**	TSC, **TY**, TSY
SPS-9	9	DMG400026428	TY, TSY
PWD	9	DMG400016613	TSC, TSY
Rca	10	DMG400019149	TSC
UGPase-11	11	DMG401013333	TSC, TY, TSY
AGPaseB-12	12	DMG400046891	TSC, TY, TSY

be associated with TSC and TY in the background of the CHIPS-ALL population (Fischer et al. *in preparation*) and with TY in the QUEST population (Chapter 2). All detected SNPs were not in the same region as the *HSP70-bad* PCR fragment.

Marker-trait associations from association mapping of candidate genes in the Quest population Seven candidate genes were selected, genotyped and tested for associations with starch and yield-related traits in Chapter 2. Six of the candidate genes were annotated in the potato genome sequence: *CP12-2*, *SSsIV*, *SssI*, *PGI1*, *Pho1b* and *BMY1*.
In total, two of the candidate loci were found to have at least one significant (FDR <0.05) SNP in any of the case-control studies. The *SssI* locus (PGSC0003DMG402018552) had 2 significant SNPs in the case-control study for TSC and 2 significant SNPs in the case-control study for TSY. The *SssI_snp6015* was associated with TY and TSY in the association mapping study (Chapter 2). All significant SNPs detected by RAD sequencing were not located in the PCR fragment that was analyzed by amplicon sequencing (Chapter 2). The *BMY1* locus (PGSC0003DMG400001855) had 1 significant SNP in the case control studies for TSC and 2 significant SNPs in the case-control study for TY. There was no evidence for any marker-trait association of the *BMY1* amplicon in the association mapping study (Chapter 2). The SNPs that were detected by RAD sequencing were not located in the PCR fragment that was analyzed by amplicon sequencing (Chapter 2).

4.3.4. Comparing the results of genotyping methods

The results of the case-control studies with RAD sequencing were compared to the results of the case-control studies with SolCAP Potato Array genotyping (Chapter 3). Annotated loci with at least one significant (FDR <0.05) SNP were compared to annotated loci that had at least one significant (p-value <0.01) SolCAP SNP (Appendix Table B.17).
The results of the comparisons are shown in Figure 4.2 (a-c). 45 loci were detected with both genotyping methods in the case-control study for TSC. 11 loci were in common between the case-control study for TY and 14 loci in the case-control study for TSY. The 65 loci (Figure 4.2 d) are listed in Table 4.5.
In the SolCAP Potato Array genotyping study (Chapter 3), six candidate loci were selected from the results of the case-control studies and tested for marker-trait associations the QUEST population by association mapping. Five of the six loci were annotated in the potato genome sequence: *F2PA*, *CP12-1*, *QUA1*, *60S* and *CIS*.
Of these five annotated loci, *CIS* (PGSC0003DMG400007797) was found in the intersecting set of the TSC and TY case-control studies (Figure 4.2 d, Table 4.5). It had 2 significant (FDR <0.05) SNPs in the case-control study for TSC and 1 significant (FDR <0.05) SNP in the case-control study for TY. With the SolCAP Potato Array genotyping, the



(a) TSC (b) TY (c) TSY

(d) intersecting set

Figure 4.2.: **Venn diagrams of the loci detected by RAD sequencing (at least one significant SNP per locus) and SolCAP Potato Array genotyping of the case-control studies for TSC (a), TY (b) and TSY (c) and the intersecting set (d).** The compared loci contained at least one significant (α=0.01) SolCAP SNP (Appendix Table B.17) or one significant (FDR <0.05) RAD sequencing SNP

Table 4.5.: **Loci detected by RAD sequencing (at least one significant SNP per locus) and SolCAP Potato Array genotyping of the case-control studies for TSC, TY or TSY or a combination of those (Figure 4.2).** In bold are Loci that were also on the compiled list of candidate genes and markers (Appendix Table B.18)

PGSC gene id (PGSC0003...)	Chromosome	Case-control study	Annotation (version v2.1.11)
DMG400006816	1	TSC	ATP binding protein
DMG400006800	1	TSC	NBS-LRR protein
DMG400027768	1	TSY	Cop11 protein
DMG400027526	1	TSY	Prolyl-tRNA synthetase
DMG400022473	1	TSC, TY	Zeta-carotene desaturase
DMG401028588	1	TSY	Nucleolar GTPase
DMG400028611	1	TSC	Peroxisomal membrane protein pmp34
DMG400030950	1	TSC	Myo inositol monophosphatase
DMG400010198	2	TSC	Conserved gene of unknown function
DMG400022953	2	TSC	Pyruvate decarboxylase
DMG400010718	2	TY	GDSL-motif lipase/hydrolase family protein
DMG401017733	2	TSC	Zinc finger protein
DMG400026392	2	TSC, TY	Mannose-6-phosphate isomerase
DMG400043061	2	TY	Serine-threonine protein kinase, plant-type
DMG400024596	3	TSC	Protein kinase atmrk1
DMG400024561	3	TSC	Ser/Thr protein kinase
DMG400000619	3	TSC	Nucleosome-binding protein
DMG400000639	3	TSC	Delta 9 desaturase
DMG400014223	3	TSC	4-coumarate–CoA ligase 2
DMG400025300	4	TSC	Conserved gene of unknown function
DMG401007955	4	TY	ALCATRAZ/SPATULA
DMG400005138	4	TSY	Glucan endo-1,3-beta-glucosidase
DMG400009981	**4**	**TSC**	**1,4-alpha-glucan branching enzyme (*SBEI*)**
DMG400010007	4	TSC	Pyrophosphate-fructose 6-phosphate 1-phosphotransferase subunit alpha
DMG400025121	5	TSC	6-phosphogluconate dehydrogenase, decarboxylating
DMG400031262	5	TSC	Methyltransferase
DMG400030978	5	TSC	Polygalacturonase non-catalytic subunit AroGP2
DMG400024449	5	TY	26S proteasome subunit
DMG400008445	5	TSC, TSY	60S ribosomal protein L6
DMG400023508	5	TSY	Kinase
DMG402007274	**6**	**TSC**	**Isoamylase isoform 3 (*DBE-6/1*)**

Table 4.5.: (continued)

PGSC gene id (PGSC0003...)	Chromosome	Case-control study	Annotation (version v2.1.11)
DMG400007297	6	TY	Protein phosphatase-2C
DMG400026598	6	TSC	Conserved gene of unknown function
DMG400027936	**7**	**TY**	**Sucrose-phosphate-synthase (*SPS-7*)**
DMG400032791	7	TY, TSY	Casein kinase
DMG400028951	7	TSC	Aconitase
DMG400009380	7	TSY	Receptor protein kinase CLAVATA1
DMG400017292	7	TSC	Protein pof4
DMG400031084	**7**	**TSC**	**ADP-glucose pyrophosphorylase, small subunit (*AGPaseB-7*)**
DMG400031099	7	TSC	Endosomal P24A protein
DMG400007070	7	TSC	Polynucleotide kinase-3'-phosphatase
DMG400022169	7	TSC	Poly(A)-specific ribonuclease PARN
DMG400014831	8	TSC	HSP DnaJ N-terminal domain-containing protein
DMG400007390	8	TSC	Beta-ketoacyl-CoA synthase
DMG400024217	8	TSC	Pseudo response regulator
DMG402012981	9	TSY	Serine-threonine protein kinase, plant-type
DMG400008917	**9**	**TY**	**Heat shock protein 70 (*HSP70*)**
DMG400029885	9	TSC	Kinase
DMG400014421	10	TSY	Proline synthetase associated protein
DMG400025001	10	TSY	Global transcription factor group
DMG400028261	10	TSC	Fructose-bisphosphate aldolase
DMG400013259	11	TSY	Nonsense-mediated mRNA decay protein
DMG400015693	11	TSC	Resistance gene
DMG400016219	11	TSC	GTP cyclohydrolase II
DMG400031071	11	TSC	Nam 9
DMG400009246	11	TSY	8-oxoguanine DNA glycosylase
DMG400030212	11	TSC	Nitrate reductase
DMG400019987	11	TSC	Tubulin-specific chaperone E
DMG400015547	11	TSY	Protein kinase
DMG400015368	12	TSC	Conserved gene of unknown function
DMG400007797	12	TSC, TY	Citrate synthase
DMG400002929	12	TSC	Chaperonin 21
DMG400001096	12	TSC	Transporter
DMG400004280	12	TSC	Phytocalpain
DMG400004277	12	TSC	Dead box ATP-dependent RNA helicase

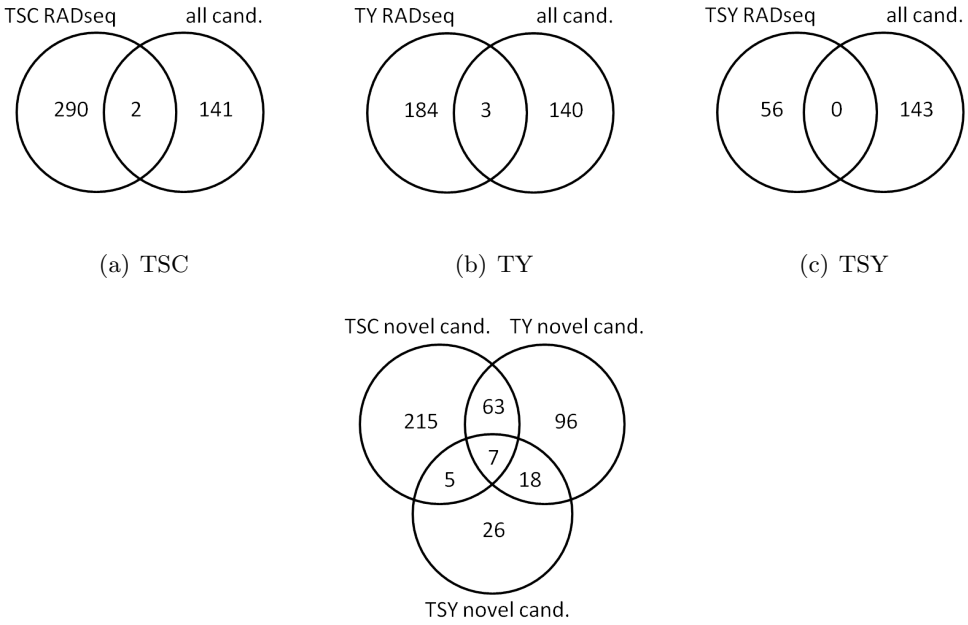
CIS_solcap_snp_c2_25372 was found to be significant in all three case-control studies (Chapter 3). There was no evidence for marker-trait associations when it was tested in all genotypes of the QUEST population. The *CIS_snp6741* was associated with TSC in the QUEST population. The significant SNPs that were detected by RAD sequencing were not located in the amplified PCR fragment for amplicon sequencing (Chapter 3).

4.3.5. Combining information of genotyping studies with compiled list of candidate genes and markers

Three loci were detected that were significant in the case-control study for TSC genotyped with RAD sequencing, in the case-control study for TSC with SolCAP Potato Array genotyping (Table 4.5) and also present on the compiled candidate gene and marker list (Appendix Table B.18): the *SBEI* locus (PGSC0003DMG400009981), the *DBE-6/1* locus (PGSC0003DMG402007274) and the *AGPaseB-7* locus (PGSC0003DMG400031084).
In the case-control study for TY, two loci were detected to be significant in the case-control studies of both genotyping methods that were present on the compiled list of candidate genes and markers: the *SPS-7* locus (PGSC0003DMG400027936) and the *HSP70* locus (PGSC0003DMG400008917), described by Fischer et al. (*in preparation*).
There was no overlap of loci with significant SNPs between the case-control study genotyped with RAD sequencing, the case-control study for TSY with SolCAP Potato Array genotyping (Chapter 3) and the compiled candidate gene list.

4.3.6. Detection of novel candidate genes by case-control studies with RAD sequencing

In order to find novel candidate gene markers by RAD sequencing of case-control studies, the results of the statistical analysis were compared to the list of candidate genes and markers (Appendix B, Table B.18). Here, only loci with at least five SNPs that had significantly (FDR <0.05) different allele frequencies between the contrasting pools, containing a minimum one SNP with non-synonymous effect on the protein, were compared to the compiled list of candidate genes and markers (see above) to remove already known candidates from the set. There were 290 novel candidate loci in the case-control study for TSC, 184 loci in the case-control study for TY and 56 loci for TSY (Figure 4.3 a-c). In total, 430 novel candidate loci were detected by the genotyping of case-control studies by RAD sequencing (Figure 4.3 d). The candidate loci were distributed over all chromosomes. They were also located in regions, where no QTL for starch or yield were described by Schäfer-Pregl et al. (1998). The novel candidate loci are depicted on the potato map (Figure 4.4) and the list of the novel candidate genes is given in Appendix B (Table B.28).



(a) TSC (b) TY (c) TSY

(d) Novel candidate genes

Figure 4.3.: **Venn diagrams of loci containing more than five significantly (FDR <0.05) different SNPs, of which minimum one had a non-synonymous effect on the protein, in comparison to the compiled list of candidate genes and markers (Appendix Table B.18).** The 430 novel loci that were not contained in the compiled list of candidate genes and markers are shown for the three case-control studies (d)

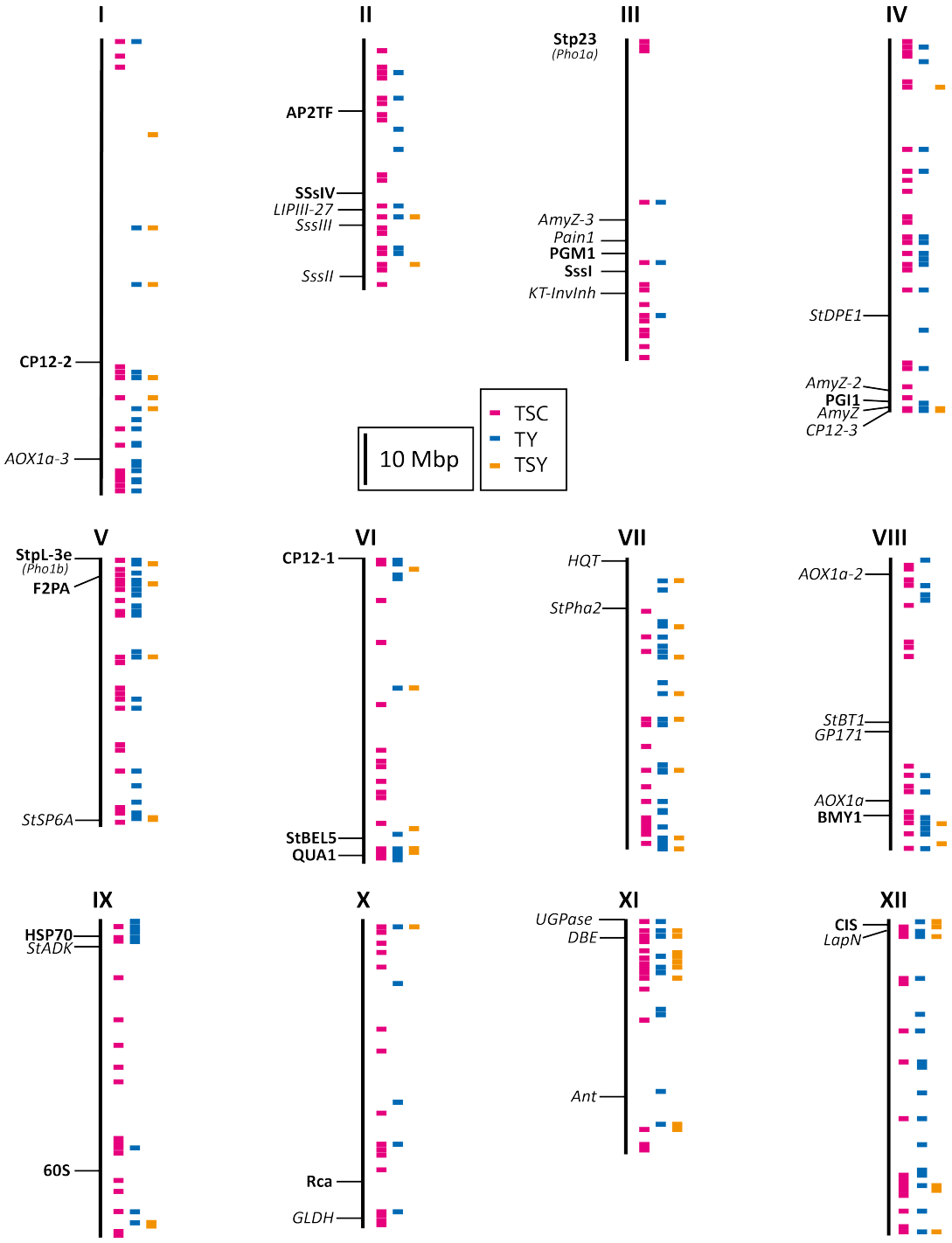
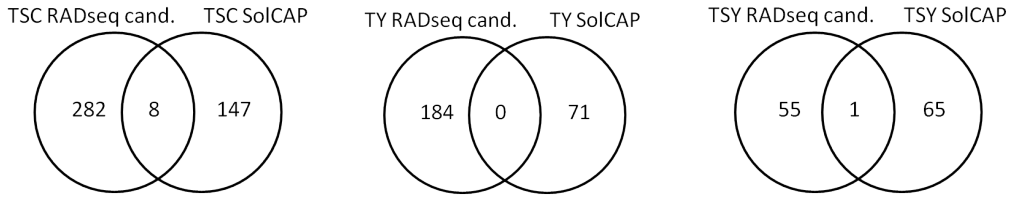


Figure 4.4.: **Physical map of potato (version v4.03).** Shown are the 430 novel candidate loci containing more than five significant (FDR <0.05) SNPs, of which minimum one had a non-synonymous effect on the protein, detected by RAD sequencing in the case-control studies for TSC, TY and TSY (Appendix Table B.28). Compiled candidates (Chapter 2) are in italics. Candidate genes and markers that were tested by association mapping (Chapters 2+3) are in bold

The set of 9 loci between the novel candidate genes from the RAD sequencing study and the loci detected by the genotyping with the SolCAP Potato Array (Chapter 3) was identified (Table 4.6). 8 loci were common between the case-control study for TSC (Figure 4.5 a). There were no common loci between the case-control studies for TY (Figure 4.5 b). One common locus was detected between the case-control studies for TSY (Figure 4.5 c). This locus was not part of the common loci detected for TSC.



(a) TSC (b) TY (c) TSY

Figure 4.5.: **Venn diagrams of novel candidate loci from RAD sequencing (Appendix Table B.28) and the SolCAP Potato Array genotyping in the case-control studies for TSC, TY and TSY (Appendix Table B.17).** The loci obtained from the SolCAP experiment contained at least one significant (α =0.01) SolCAP SNP (Chapter 3)

Table 4.6.: **Novel candidate loci from RAD sequencing (Appendix Table B.28) and the SolCAP Potato Array genotyping in the case-control studies for TSC, TY and TSY.** The loci obtained from the SolCAP experiment contained at least one significant (α=0.01) SolCAP SNP (Chapter 3)

PGSC gene id (PGSC0003...)	Chromosome	Case-control study	Annotation (version v2.1.11)
DMG400006800	1	TSC	NBS-LRR protein
DMG400022473	1	TSC	Zeta-carotene desaturase
DMG400010198	2	TSC	Conserved gene of unknown function
DMG400031262	5	TSC	Methyltransferase
DMG400030978	5	TSC	Polygalacturonase non-catalytic subunit AroGP2
DMG400023508	5	TSY	Kinase
DMG400007390	8	TSC	Beta-ketoacyl-CoA synthase
DMG400031071	11	TSC	Nam 9
DMG400002929	12	TSC	Chaperonin 21

4.4. Discussion

4.4.1. Novel candidate genes detected for starch-yield traits by RAD sequencing in case-control studies

Novel candidate genes for TSC, TY and TSY were detected by RAD sequencing in a case-control study design. In total, 430 novel and non-obvious candidate genes were detected in the case-control studies. The candidate loci were distributed over all chromosomes and also located in regions with no QTL for starch or yield described (Schäfer-Pregl et al., 1998). There were two major selection criteria for the detection of novel candidate genes. Firstly, only non-synonymous SNPs were taken into account for the selection, as the non-synonymous effect on the protein may lead to a change in the efficiency of the protein, the binding affinity or the formation of complexes. Thereby, a difference in phenotypic variation may be explained. Secondly, more than five significant SNPs had to be present in the locus. As a result, the number of false-positives was expected to be reduced as a higher number of significant SNPs was suggested to ensure a region that is associated with the relevant trait. The nine non-obvious candidate genes, that were detected in all three case-control studies by RAD sequencing as well as by SolCAP Potato Array genotyping (Chapter 3), are of special interest for further analysis. Further steps could be the validation of the associations in a larger genetic background by the design and testing of allele-specific primers for the highly significant SNPs of the novel candidates. SNPs with validated marker-trait associations can then directly be applied by marker-assisted selection in potato breeding programs. Still, the discovered candidates may not be the causal genes for the studied traits and the functional analysis of candidate genes could provide further information.

4.4.2. RAD sequencing valuable tool for genome-wide genotyping of tetraploid potato

RAD sequencing is a valuable tool for genome-wide genotyping of tetraploid potatoes. This is the first study of genotyping by RAD sequencing in potato. More than half a million SNPs, distributed over all chromosomes, were detected in each case-control study. Depending on the assumptions of SNP frequency in potato, the calculated number of expected SNPs ranged between less than 200,000 (1 SNP/140 bp in exon; Uitdewilligen, 2012) and more than 1 million (1 SNP/21 bp in exon; Rickert et al., 2003). Around 300,000 SNPs were expected at a frequency of 1 SNP/80 bp, which was reported for intron regions (Uitdewilligen, 2012). The RAD sequences in this study were not only generated from exon and intron regions, but also from non-coding regions of the genome. The number of detected SNPs lay above the expected number at a SNP frequency of 1 SNP/80 bp, which

may be due to a higher nucleotide diversity in the non-coding regions of the genome. The extensive number of more than half a million SNPs in each case-control study allowed for a genome-wide statistical analysis.

4.4.3. Marker-trait associations from previous studies confirmed by RAD sequencing

A number of marker-trait associations with TSC, TY and TSY, reported in previous studies, were validated by genotyping by RAD sequencing and case-control analysis. Two loci were detected with the case-control analysis, containing PCR markers that were associated with starch content and yield in previous studies. The *Rca* locus was detected in the case-control study for TSC, confirming the previous association of the *Rca1a* allele (Li et al., 2008) with TSC (Chapter 2). Similarly, the associations of the *HSP70-bad* allele (Fischer et al. *in preparation*) with TSC and TY were confirmed. Due to pooling the sequences prior to SNP analysis, the co-segregation of RAD sequencing SNPs with the allele specific markers could not be tested. Furthermore, marker-trait associations of candidate genes, which were tested in the genetic background of the QUEST population (Chapter 2+3), were confirmed in the case-control studies with RAD sequencing. The association of the *SssI* locus (Chapter 2) with TSY was confirmed as well as the association of *CIS* with TSC (Chapter 3). Thus, previously reported marker-trait associations in the same as well as different populations could be verified by the case-control study design, confirming the practicability of RAD sequencing for the detection of candidate genes.
However, not all reported marker-trait associations in the QUEST population were significant in the case-control study. There are several explanations for the lack of associations in the case-control studies. Firstly, the locus may not have been represented in the generated sequences. The genomic sequence of a candidate locus has to contain a minimum of one restriction site in order to be represented by a RADtag. Secondly, there may have been a restriction site polymorphism. This would lead to a null-allele which does not allow for the detection of all haplotypes and therefore the true variation of the locus would be missing. Thirdly, the sequence variation detected by RAD sequencing was too distal from the previously associated polymorphism. If the LD decay was high between the two regions within the candidate gene, it might be that the marker-trait association could not be detected. Furthermore, SNPs were selected based on location in annotated genes and *Pho1a* SNPs, for example, are not detected due to this approach. Therefore, the reasons for the lack of previous associations can be manifold.
For one tested candidate gene (*BMY1*) there was no evidence for marker-trait associations in the entire QUEST population (Chapter 2). In contrast to that, significant SNPs were detected in the case-control studies. The significant SNPs from the case-control studies

were not located in the region of the previously tested amplicon. The tested amplicon for *BMY1* had furthermore a length of less than 350 bp, covering 10 SNPs. A possible explanation for the detection of significant SNPs in the case-control study, contrary to expectations, may be that the amount of SNPs was not sufficient to tag all existing haplotypes. The newly detected significant SNPs might tag different haplotypes. However, this could not be tested as the sequences were pooled prior to the analysis and no genotype data of individuals was available. A re-analysis of sequence data would be required for the test of co-segregation between RAD sequencing SNPs and amplicon sequencing SNPs. The design of new allele-specific primers based on the sequencing information may lead to the detection of novel marker-trait associations in the full QUEST population.

4.4.4. Marker-trait associations detected for SNPs in knowledge-based candidate genes

Marker-trait associations in knowledge-based candidate genes were detected by RAD sequencing. All loci with significant SNPs in the case-control studies were compared to a compiled list of candidate genes. The list contained (i) starch- and yield-related candidate genes from this thesis (Chapter 2) and (ii) candidate genes operational in starch sugar interconversion (Schreiber et al. *in preparation*). The four top hits, which were also found in the case-control study with SolCAP Potato Array genotyping were the associations of *SBEI*, *DBE-6/1* and *AGPaseB-7* with TSC and the association of *SPS-7* with TY. In general, the detection of significant SNPs in the knowledge-based candidate loci illustrates the potential of the case-control studies with RAD sequencing for the discovery of marker-trait associations. In the case of the knowledge-based candidate loci, allele-specific markers can be developed from the sequencing information for the directly implementation in marker-assisted selection.

4.4.5. Limitations of RAD sequencing in tetraploid potatoes

The RAD sequencing method reduces the complexity of the genome. Therefore, it allows for a higher number of multiplexing on sequence lanes, which reduces the cost of high-density SNP detection. Still, there are two main limitations of the method, that need to be considered. The first limitation is that by the reduction of complexity, not all loci are represented by the sequences generated. In order to increase the amount of coding regions that are sequenced, the libraries could be prepared based on transcriptome sequence. Still, it would be unlikely that all loci were represented due to the tissue- and time point-specific expression of genes. Another possibility to influence the amount of coding regions that are sequenced is the choice of restriction enzyme. A restriction enzyme, which selectively cuts in coding regions of the genome, might be desired. However, new

restriction enzymes require the design of new individual P1 adapters, which is the major matter of expense in the experiment, aside from the sequencing. In this study, the *Kpn*I restriction enzyme was chosen. It is a rare cutting enzyme that cuts in GC rich regions. Thereby, a desired bias towards coding regions was introduced. A second limitation of the RAD sequencing method is the high heterozygosity of tetraploid potatoes. With the large number of polymorphisms, the chance for a restriction site polymorphism is high. The polymorphism in a restriction site leads to a null allele because the haplotypes lacking the restriction site are not detected. Keeping these limitations in mind, RAD sequencing is a valuable tool for genome-wide genotyping in tetraploid potatoes.

4.4.6. Concluding paragraph

Genome-wide SNP discovery by RAD sequencing in tetraploid potato was applied for the first time in a case-control study design with the objective to find novel candidate genes for complex traits. In total 430 novel non-obvious candidate genes were discovered for starch and yield-related traits. Furthermore, new marker-trait associations with knowledge-based candidate genes were detected. Based on the sequence information, markers for direct implementation in marker-assisted selection can be generated. The results of this study show that RAD sequencing in a case-control study design is an efficient tool for the detection of novel candidate genes for complex traits in polyploid species.

5. General discussion

This thesis focused on the detection of novel SNP markers and candidate genes for potato tuber starch and yield-related traits, such as tuber starch content and tuber yield, in the framework of the project "Quality Starches by Exploiting New Breeding Tools in *Solanum tuberosum*" (QUEST). In order to pursue this research objective, different genotyping methods were applied in combination with association mapping and case-control study design.

5.1. Novel markers and candidate genes for starch yield traits by combined association mapping and case-control analysis lead to

A novel set of markers and candidate genes was detected for starch yield, tuber starch content and tuber yield. Combining association mapping and a case-control design for the detection of sequence variation associated with phenotypic variation lead to the discovery of novel SNP markers and candidate genes in tetraploid potatoes. Both approaches make use of assembled populations (Table 5.1). The individuals are related by descent and share historic meiotic recombination events (Flint-Garcia et al., 2003). A lower number of genotypes is required for a case-control study than for association mapping. Marker-trait associations by association mapping are detected by the implementation of rather challenging statistics - a mixed linear model including population structure and kinship between the genotypes - while the statistical analysis of case-control studies is much more straightforward. The result of the case-control study is a pre-selection of genetic markers. These markers need further confirmation in a larger population size as loci identified in smaller population sizes are not always significant in a larger genetic background (Huang et al., 2011). The markers obtained from association mapping, on the other hand, can be directly implemented in breeding programs for marker-assisted selection (Flint-Garcia et al., 2003). Both approaches have their special strengths and limitations. This thesis showed the value of combining the approaches in order to find novel SNP markers and candidate genes for complex traits in potato.

Table 5.1.: **Comparison of association mapping and case-control study design for the detection of marker-trait associations**

Features	Association mapping	Case-control study
Number of genotypes	high	low
Implementation of results	direct application in breeding programs	needs further confirmation
Statistics	difficult	easy

5.2. Candidate gene based and high-density genotyping methods have individual strengths and limitations

Three genotyping methods were applied for the detection of genotypic variation in the genetic background of the QUEST population. A lower-density genotyping method was used in association mapping and two high-density genotyping methods were tested in a case-control study design. With Sanger amplicon sequencing, the individuals were genotyped for a small number of knowledge-based candidate genes. The PCR fragments were designed so they yielded between 10 and 40 SNPs (Table 5.2). The genotypes were analyzed with the SolCAP Potato Array for 8,303 SNPs in each individual. By RAD sequencing, more than half a million SNP positions were detected in the samples. However, the difficulties with sample preparation and data processing increased with the number of SNPs that were detected. Similarly, the cost were comparably high for the genotyping by RAD sequencing compared to Sanger amplicon sequencing and the SolCAP Potato Array genotyping. The main advantage of the genome-wide high-density genotyping methods was the detection of novel candidate genes for the traits of interest, while Sanger amplicon sequencing relied on previous knowledge about the genes. All genotyping methods allowed for the detection of sequence variation in potato individuals. Variation in knowledge-based candidate genes was detected by the Sanger amplicon sequencing method, while SNPs in non-obvious candidate loci could be detected by genotyping with the SolCAP Potato Array and RAD sequencing.

Is it more desirable to genotype potatoes with an array or with next-generation sequencing technologies? Genotyping with the SolCAP Potato Array generates information about a maximum of 8,303 SNPs of each individual. However, the method had two shortcomings. Firstly, the SNP loci on the array were selected based on their function and relevance (Felcher et al., 2012). Secondly, the SNP information was obtained mainly based on North American potato cultivars. Thereby it is missing valuable alleles of characteristic European potato alleles (Chapter 3). Uitdewilligen (2012) developed a potato array containing 20,000 SNP markers. The SNPs were selected based on sequencing information of 80

Table 5.2.: **Comparing genotyping methods for the detection of sequence variants in potato.** The genotyping was performed in the QUEST population (Chapters 2 to 4)

Features	Sanger amplicon sequencing	SolCAP Potato Array	RAD sequencing
SNP detection	10-40/fragment	≤8,303	<500,000
Sample preparation	easy	easy	difficult
Data processing	easy	medium	difficult
Cost	low	low	high

European potato cultivars and the array was designed to include rare alleles as well. However, this array is also knowledge-based. In order to find novel candidate loci for complex traits it is desired to genotype with an unbiased approach. In this thesis, the applicability of RAD sequencing for genotyping tetraploid potato cultivars was tested. The approach is rather unbiased, although the detection of sequencing variants depend on the presence of the recognition site of the restriction enzyme. With the decreasing cost of next-generation sequencing, it will become more feasible to apply large-scale genotyping methods that are based on this technique. These methods will eventually outperform the array technology, once analysis-pipelines become more user-friendly.

5.3. Sketching a picture of potato breeding in the future

The future of marker-trait association studies in potato breeding programs is likely to be situated in a more quantitative marker-assisted selection system. At present, potato breeding is still a matter of the "breeder's eye", with the support of a limited amount of molecular markers. In other crops, such as corn and wheat, the concept of genomic selection is increasingly applied (Crossa et al., 2013). Genomic selection is a strategy, where a diverse population of individuals (training population) is intensively genotyped and phenotyped in several years and locations. This training population is then used for the development of marker sets that can predict the field performance of individuals based on their genotypes (Cabrera-Bosquet et al., 2012). The success of this method strongly depends on the quality of the training population. Trend-setting would be the establishment of a potato collection as training population with diverse germplasm, phenotyped for desired traits in world-wide potato growing regions. Genome-wide high-density genotyping should be performed with a next-generation sequencing method, such as RAD sequencing (Chapter 4) or a genotyping-by-sequencing approach (e.g. Uitdewilligen et al., 2013). The most challenging part is most likely the phenotyping over many years at different locations. Firstly, the quarantine regulations for shipping plant material between countries

are rather strict due to regulations on sanitary conditions. This may be a problem for the world-wide distribution of potato tubers. Secondly, the phenotyping methods have to be standardized across all locations and years. The breeders and growers will have to provide space for the repeated planting of the training population and they need to invest the time for the phenotyping. Another major challenge will be the data processing. However, a successful example for data analysis for genomic prediction on a public database is already available for casava (Casavabase; http://www.cassavabase.org/breeders/index.pl) and is upcoming for tomato (Sol Genomics Network; http://solgenomics.net/breeders/index.pl). There is a realistic chance for the successful application of genomic selection for complex traits in potato in the future, if these challenges can be overcome.

5.4. Concluding remarks

In this thesis, three concepts were applied with the objective to detect sequence variation that explains phenotypic variation in starch and yield-related traits in potato: association mapping with a knowledge-based candidate gene approach (Chapter 2), a case-control study with genotyping by the SolCAP Potato Array (Chapter 3) and genotyping with next-generation RAD sequencing (Chapter 4). All three concepts resulted in the detection of novel marker-trait associations and candidate genes. Association mapping of knowledge-based candidate genes resulted in the detection of 14 diagnostic SNP markers and one indel. The case-control studies for tuber starch content, tuber yield and starch yield in combination with SolCAP Potato Array genotyping lead to the detection of 328 marker-trait associations, of which seven SNP markers were validated by association mapping in the full population. By genotyping with next-generation RAD sequencing of the case-control studies, 430 novel candidate genes were discovered for tuber starch content, tuber yield and starch yield. Further confirmation of the novel candidate genes will be obtained by the genotyping and analysis of the respective polymorphisms in a larger genetic background. The results of this thesis show that the genome-wide discovery of polymorphisms is an efficient tool for the detection of novel candidate genes and markers. The decreasing cost for next-generation sequencing will allow for the genotyping of an entire association mapping population by RAD sequencing. For further studies, this genotyping method might be of high value in combination with association mapping.

References

Abel, G. J., Springer, F., Willmitzer, L., and Kossmann, J. (1996). Cloning and functional analysis of a cDNA encoding a novel 139 kDa starch synthase from potato (*Solanum tuberosum* L.). *The Plant Journal*, 10(6):981–991.

Anithakumari, A. M., Dolstra, O., Vosman, B., Visser, R. G. F., and Van der Linden, C. G. (2011). In vitro screening and QTL analysis for drought tolerance in diploid potato. *Euphytica*, 181(3):357–369.

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P., and Nordborg, M. (2005). Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLOS Genetics*, 1:e60.

Avebe (2013). http://www.avebe.com/ (17-11-2013).

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE*, 3(10):7.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–91.

Ballvora, A., Flath, K., Lübeck, J., Strahwald, J., Tacke, E., Hofferbert, H.-R., and Gebhardt, C. (2011). Multiple alleles for resistance and susceptibility modulate the defense response in the interaction of tetraploid potato (*Solanum tuberosum*) with *Synchytrium endobioticum* pathotypes 1, 2, 6 and 18. *Theoretical and Applied Genetics*, 123(8):1281–92.

Barchi, L., Lanteri, S., Portis, E., Acquadro, A., Vale, G., Toppino, L., and Rotino, G. L. (2011). Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics*, 12(1):304.

Becker, K. and Leithold, G. (2008). Die Verbesserung des Vorfruchtwertes von Winterweizen durch den Einsatz von legumen Untersaaten im Anbauverfahren Weite Reihe. pages 1–2. Gesellschaft für Pflanzenbauwissenschaften.

Bonierbale, M. W., Plaisted, R. L., and Tanksley, S. D. (1988). RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. *Genetics*, 120:1095–1103.

Botstein, D., White, R. L., and Skolnick, M. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*.

Bradshaw, J. E., Hackett, C. a., Meyer, R. C., Milbourne, D., McNicol, J. W., Phillips, M. S., and Waugh, R. (1998). Identification of AFLP and SSR markers associated with quantitative resistance to *Globodera pallida* (Stone) in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*) with a view to marker-assisted selection. *Theoretical and applied genetics*, 97(1-2):202–210.

Bradshaw, J. E., Hackett, C. A., Pande, B., Waugh, R., and Bryan, G. J. (2008). QTL mapping of yield, agronomic and quality traits in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*). *Theoretical and Applied Genetics*, 116(2):193–211.

Breseghello, F. and Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics*, 172(2):1165–77.

Bundesministerium für Ernährung Landwirtschaft und Verbraucherschutz (2013). Kartoffelernte 2013 liegt weit unter dem mehrjährigen Durchschnitt. *Pressemitteilung Nr. 268 vom 26.09.13*.

Bus, A., Hecht, J., and Huettel, B. (2012). High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *BMC Genomics*.

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., and Araus, J. L. (2012). High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. *Journal of integrative plant biology*, 54(5):312–20.

Chen, H., Rosin, F. M., Prat, S., and Hannapel, D. J. (2003). Interacting transcription factors from the three-amino acid loop extension superclass regulate tuber formation. *Plant Physiology*, 132(3):1391–1404.

Chen, X., Hedley, P. E., Morris, J., Liu, H., Niks, R. E., and Waugh, R. (2011). Combining genetical genomics and bulked segregant analysis-based differential expression: an approach to gene localization. *Theoretical and Applied Genetics*, 122(7):1375–83.

Chen, X., Salamini, F., and Gebhardt, C. (2001). A potato molecular-function map for carbohydrate metabolism and transport. *Theoretical and Applied Genetics*, 102(2):284–295.

Chutimanitsakun, Y., Nipper, R. W., Cuesta-Marcos, A., Cistué, L., Corey, A., Filichkina, T., Johnson, E. A., and Hayes, P. M. (2011). Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC Genomics*, 12(1):4.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Ruden, D. M., and Lu, X. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w1118; iso-2; iso-3*. *Fly*, 6(2):1–13.

Cockram, J., White, J., Leigh, F. J., Lea, V. J., Chiapparino, E., Laurie, D. A., Mackay, I. J., Powell, W., and O'Sullivan, D. M. (2008). Association mapping of partitioning loci in barley. *BMC genetics*, 9:16.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., and Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, 142(1-2):169–196.

Crossa, J., Perez, P., Hickey, J., Burgueno, J., Ornella, L., Ceron-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., and Mathews, K. (2013). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, (January):1–13.

D'hoop, B. B., Paulo, M.-J., Kowitwanich, K., Sengers, M., Visser, R. G. F., Van Eck, H. J., and Van Eeuwijk, F. (2010). Population structure and linkage disequilibrium unravelled in tetraploid potato. *Theoretical and Applied Genetics*, 121(6):1151–1170.

D'hoop, B. B., Paulo, M.-J., Mank, R., Van Eck, H. J., and Van Eeuwijk, F. (2008). Association mapping of quality traits in potato (*Solanum tuberosum* L.). *Euphytica*, 161(1):47–60.

Donald, T. M., Pellerone, F., Adam-Blondon, A.-F., Bouquet, A., Thomas, M. R., and Dry, I. B. (2002). Identification of resistance gene analogs linked to a powdery mildew resistance locus in grapevine. *Theoretical and Applied Genetics*, 104(4):610–618.

Douches, D. S. and Freyre, R. (1994). Identification of genetic factors influencing chip color in diploid potato (*Solanum spp.*). *American Potato Journal*, 71:581–590.

Draffehn, A. M., Meller, S., Li, L., and Gebhardt, C. (2010). Natural diversity of potato (*Solanum tuberosum*) invertases. *BMC Plant Biology*, 10(271):271.

Earl, D. A. and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4(2):359–361.

Ellis, R. P., Cochrane, M. P., Dale, M. F. B., Duffus, C. M., Lynn, A., Morrison, I. M., Prentice, R. D. M., Swanston, J. S., and Tiller, S. A. (1998). Starch production and industrial use. *Journal of the Science of Food and Agriculture*, 77(3):289–311.

Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., and Cresko, W. A. (2011). SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. In Orgogozo, V. and Rockman, M. V., editors, *Methods*, volume 772 of *Methods in Molecular Biology*, pages 157–178. Humana Press, Totowa, NJ.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14(8):2611–20.

Fachverband der Stärke-Industrie (2013). Fachverband der Stärke-Instustrie e.V. *Zahlen und Daten zur deutschen Stärke-Industrie*. http://www.staerkeverband.de/html/zahlen.html (18-11-2013).

FAO Crops Statistics Database (2013). FAOSTAT Database on Agriculture. http://faostat.fao.org/ (17-11-2013).

Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., Buell, C. R., and Douches, D. S. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLOS ONE*, 7(4):e36347.

Fischer, M., Schreiber, L., Colby, T., Kuckenberg, M., Tacke, E., Hofferbert, H.-R., Schmidt, J., and Gebhardt, C. (2013). Novel candidate genes influencing natural variation in potato tuber cold sweetening identified by comparative proteomics and association mapping. *BMC Plant Biology*, 13:113.

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology*, 54(1):357–374.

Freyre, R. and Douches, D. S. (1994). Development of a model for marker-assisted selection of specific gravity in diploid potato across environments. *Crop Science*, 34:1361–1368.

Gebhardt, C. (2005). Potato genetics: molecular maps and more. In Lörz, H. and Wenzel, G., editors, *Biotechnology in Agriculture and Forestry*, volume 55, chapter II, pages 215–227. Springer, Berlin Heidelberg.

Gebhardt, C. (2007). Molecular markers, maps and population genetics. In Vreugdenhil, D., Bradshaw, J., Gebhardt, C., Govers, F., Taylor, M. A., MacKerron, D. K., and Ross, H. A., editors, *Potato Biology and Biotechnology: Advances and Perspectives*, chapter 5, pages 77–89. Elsevier, London.

Gebhardt, C. (2013). Bridging the gap between genome analysis and precision breeding in potato. *Trends in Genetics*, 29(4):248–56.

Gebhardt, C., Ballvora, A., Walkemeier, B., Oberhagemann, P., and Schuler, K. (2004). Assessing genetic potential in germplasm collections of crop plants by marker-trait association: a case study for potatoes with quantitative variation of resistance to late blight and maturity type. *Molecular Breeding*, 13(1):93–102.

Gebhardt, C., Menéndez, C. M., Chen, X., Li, L., Schäfer-Pregl, R., and Salamini, F. (2005). Genomic approaches for the improvement of tuber quality traits in potato. *Acta Horticulturae (ISHS)*, 684:85–92.

Gebhardt, C., Ritter, E., Barone, A., Debener, T., Walkemeier, B., Schachtschabel, U., Kaufmann, H., Thompson, R. D., Bonierbale, M. W., Ganal, M. W., Tanksley, S. D., and Salamini, F. (1991). RFLP maps of potato and their alignment with the homoeologous tomato genome. *Theoretical and Applied Genetics*, 83(1):49–57.

Gebhardt, C., Ritter, E., Debener, T., Schachtschabel, U., Walkemeier, B., Uhrig, H., and Salamini, F. (1989). RFLP analysis and linkage mapping in *Solanum tuberosum*. *Theoretical and Applied Genetics*, 78(1):65–75.

Gebhardt, C. and Valkonen, J. P. (2001). Organization of genes controlling disease resistance in the potato genome. *Annual review of phytopathology*, 39:79–102.

Ghislain, M., Nunez, J., Del Rosario Herrera, M., Pignataro, J., Guzman, F., Bonierbale, M. W., and Spooner, D. M. (2009). Robust and highly informative microsatellite-based genetic identity kit for potato. *Molecular Breeding*, 23(3):377–388.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–338.

Griffiths, A. J. F., Wessler, S. R., Lewontin, R. C., Gelbart, W. M., Suzuki, D. T., and Miller, J. H. (2005). *Introduction to Genetic Analysis*. W. H. Freeman and Company, New York, 8th edition.

Hackett, C. A., McLean, K., and Bryan, G. J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PLOS ONE*, 8(5):e63939.

Hamilton, J. P., Hansey, C. N., Whitty, B. R., Stoffel, K., Massa, A. N., Van Deynze, A., De Jong, W. S., Douches, D. S., and Buell, C. R. (2011). Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genomics*, 12(1):302.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hosaka, K. and Sanetomo, R. (2012). Development of a rapid identification method for potato cytoplasm and its use for evaluating Japanese collections. *Theoretical and Applied Genetics*, 125(6):1237–51.

Hovenkamp-Hermelink, J. H. M., De Vries, J. N., Adamse, P., Jacobsen, E., Witholt, B., and Feenstra, W. J. (1988). Rapid estimation of the amylose/amylopectin ratio in small amounts of tuber and leaf tissue of the potato. *Potato Research*, 31(2):241–246.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C., Fan, D., Lu, Y., Weng, Q., Liu, K., Zhou, T., Jing, Y., Si, L., Dong, G., Huang, T., Lu, T., Feng, Q., Quian, Q., Li, J., and Han, B. (2011). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, 44:32–39.

Jansen, G., Flamme, W., Schüler, K., and Vandrey, M. (2001). Tuber and starch quality of wild and cultivated potato species and cultivars. *Potato Research*, 44:137–146.

Jobling, S. A. (2004). Improving starch for food and industrial applications. *Current Opinion in Plant Biology*, 7(2):210–218.

Jonik, C., Sonnewald, U., Hajirezaei, M.-R., Flügge, U.-I., and Ludewig, F. (2012). Simultaneous boosting of source and sink capacities doubles tuber starch yield of potato plants. *Plant Biotechnology Journal*, 10(9):1088–98.

Kang, H. M., Zaitlen, N. a., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23.

Kloosterman, B., Oortwijn, M., Uitdewilligen, J., America, T., De Vos, R., Visser, R. G. F., and Bachem, C. W. B. (2010). From QTL to candidate gene: genetical genomics of simple and complex traits in potato using a pooling strategy. *BMC genomics*, 11:158.

Kossmann, J., Emmermann, M., Virgin, I., and Renz, A. (1996). DNA molecules coding for debranching enzymes derived from plants.

Kreike, C. M., De Koning, J. R., Vinke, J. H., Van Ooijen, J. W., and Stiekema, W. J. (1994). Quantitatively-inherited resistance to *Globodera pallida* is dominated by one major locus in *Solanum spegazzinii*. *Theoretical and Applied Genetics*, 88(6-7):764–9.

Krusiewicz, D., Jakuczun, H., Wasilewicz-Flis, I., Strzelczyk-Zyta, D., and Marczewski, W. (2011). Molecular mapping of the *AOX1a* and *β-AmyI* genes in potato. *Plant Breeding*, 130(4):500–502.

Lado, B., Matus, I., Rodriguez, A., Inostroza, L., Poland, J., Belzile, F., Del Pozo, A., Quincke, M., and von Zitzewitz, J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes, Genomes, Genetics*.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

Leroch, M., Kirchberger, S., Haferkamp, I., Wahl, M., Neuhaus, H. E., and Tjaden, J. (2005). Identification and characterization of a novel plastidic adenine nucleotide uniporter from *Solanum tuberosum*. *Journal of Biological Chemistry*, 280(18):17992–18000.

Li, L., Paulo, M.-J., Strahwald, J., Lübeck, J., Hofferbert, H.-R., Tacke, E., Junghans, H., Wunder, J., Draffehn, A., Van Eeuwijk, F., and 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(8):1167–81.

Li, L., Paulo, M.-J., Van Eeuwijk, F., and 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(7):1303–1310.

Li, L., Strahwald, J., Hofferbert, H.-R., Lubeck, J., Tacke, E., Junghans, H., Wunder, J., and Gebhardt, C. (2005). DNA variation at the invertase locus *invGE/GF* is associated with tuber quality traits in populations of potato breeding clones. *Genetics*, 170(2):813–821.

Li, L., Tacke, E., Hofferbert, H.-R., Lübeck, J., Strahwald, J., Draffehn, A. M., Walkemeier, B., and Gebhardt, C. (2013). Validation of candidate gene markers for marker-assisted selection of potato cultivars with improved tuber quality. *Theoretical and Applied Genetics*, 126(4):1039–52.

Li, X., Van Eck, H. J., Rouppe van der Voort, J. N. A. M., Huigen, D.-J., Stam, P., and 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(8):1121–1128.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–9.

Lloyd, J., Blennow, A., Burhenne, K., and Kossmann, J. (2004). Repression of a novel isoform of disproportionating enzyme (stDPE2) in potato leads to inhibition of starch degradation in leaves but not tubers stored at low temperature. *Plant Physiology*, 134(4):1347.

Loiselle, B. A., Sork, V. L., Nason, J., and Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, 82(11):1420–1425.

Lu, Y. A. N. and Sharkey, T. D. (2006). The importance of maltose in transitory starch breakdown. *Plant, Cell & Environment*, 29:353–366.

Mahl, J. (2013). Südstärke GmbH, personal communication 18-11-2013.

Malosetti, M., Van der Linden, C. G., Vosman, B., and Van Eeuwijk, F. A. (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. *Genetics*, 175(2):879–89.

Max Planck (1932). *Where is science going?* W.W. Norton & Company, Inc., New York.

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

Meksem, K., Leister, D., Peleman, J., Zabeau, M., Salamini, F., and Gebhardt, C. (1995). A high-resolution map of the vicinity of the *R1* locus on chromosome V of potato based on RFLP and AFLP markers. *Molecular and General Genetics*, 249(1):74–81.

Menéndez, C. M., Ritter, E., Schäfer-Pregl, R., Walkemeier, B., Kalde, A., Salamini, F., and Gebhardt, C. (2002). Cold sweetening in diploid potato: mapping quantitative trait loci and candidate genes. *Genetics*, 162(3):1423–34.

Meyer, R. C., Milbourne, D., Hackett, C. A., Bradshaw, J. E., McNichol, J. W., and Waugh, R. (1998). Linkage analysis in tetraploid potato and association of markers

with quantitative resistance to late blight (*Phytophthora infestans*). *Molecular and General Genetics*, 259:150–160.

Milbourne, D., Meyer, R. C., Collins, A. J., Ramsay, L. D., Gebhardt, C., and Waugh, R. (1998). Isolation, characterisation and mapping of simple sequence repeat loci in potato. *Molecular and General Genetics*, 259(3):233–245.

Milbourne, D., Pande, B., and Bryan, G. J. (2007). Potato. In Kole, C., editor, *Genome Mapping and Molecular Breeding in Plants. Pulses, Sugar and Tuber Crops*, volume 3, chapter 12, pages 205–236. Springer, Berlin Heidelberg.

Muth, J., Hartje, S., Twyman, R. M., Hofferbert, H.-R., Tacke, E., and Prüfer, D. (2008). Precision breeding for novel starch variants in potato. *Plant Biotechnology Journal*, 6(6):576–84.

Navarro, C., Abelenda, J. A., Cruz-Oró, E., Cuéllar, C. A., Tamaki, S., Silva, J., Shimamoto, K., and Prat, S. (2011). Control of flowering and storage organ formation in potato by FLOWERING LOCUS T. *Nature*, 478(7367):119–22.

Ogden, R., Gharbi, K., Mugue, N., Martinsohn, J., Senn, H., Davey, J. W., Pourkazemi, M., McEwing, R., Eland, C., Vidotto, M., Sergeev, A., and Congiu, L. (2013). Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology*, 22(11):3112–23.

Oliemans, W. H. (1988). *Het brood van de armen: de geschiedenis van de aardappel temidden van ketters, kloosterlingen en kerkvorsten*. SDU uitgeverij, 's Gravenhage.

Ortega, F. and Lopez-Vizcon, C. (2012). Application of molecular marker-assisted selection (MAS) for disease resistance in a practical potato breeding programme. *Potato Research*, 55(1):1–13.

Pajerowska-Mukhtar, K. M., Stich, B., Achenbach, U., Ballvora, A., Lubeck, J., Strahwald, J., Tacke, E., Hofferbert, H.-R., Ilarionova, E., Bellin, D., Walkemeier, B., Basekow, R., Kersten, B., and Gebhardt, C. (2009). Single nucleotide polymorphisms in the *allene oxide synthase 2* gene are associated with field resistance to late blight in populations of tetraploid potato cultivars. *Genetics*, 181(3):1115–1127.

Petersen, R. (1985). Augmented designs for preliminary yield trials (revised). *RACHIS (ICARDA); Barley, Wheat and Triticale Newsletter*, 4(1):27–32.

Pfender, W. F., Saha, M. C., Johnson, E. A., and Slabaugh, M. B. (2011). Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theoretical and Applied Genetics*, 122:1467–1480.

Pohlmeyer, K., Paap, B. K., Soll, J., and Wedel, N. (1996). CP12: a small nuclear-encoded chloroplast protein provides novel insights into higher-plant GAPDH evolution. *Plant Molecular Biology*, 32(5):969–78.

Pritchard, J. K. and Donnelly, P. (2001). Case-control studies of association in structured or admixed populations. *Theoretical Population Biology*, 60(3):227–37.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Provan, J., Powell, W., and Waugh, R. (1996). Microsatellite analysis of relationships within cultivated potato (*Solanum tuberosum*). *Theoretical and Applied Genetics*, 92(8):1078–1084.

Quarrie, S. A., Lazic-Jancic, V., Kovacevic, D., Steed, A., and Pekic, S. (1999). Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. *Journal of Experimental Botany*, 50(337):1299–1306.

R Development Core Team (2013). R: A Language and Environment for Statistical Computing.

Regierer, B., Fernie, A. R., Springer, F., Perez-Melis, A., Leisse, A., Koehl, K., Willmitzer, L., Geigenberger, P., and Kossmann, J. (2002). Starch content and yield increase as a result of altering adenylate pools in transgenic plants. *Nature Biotechnology*, 20(12):1256–60.

Reif, J. C., Melchinger, A. E., and Frisch, M. (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science*, 45(1):1–7.

Rickert, A. M., Kim, J. H., Meyer, S., Nagel, A., Ballvora, A., Oefner, P. J., and Gebhardt, C. (2003). First-generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. *Plant Biotechnology Journal*, 1(6):399–410.

Rickert, A. M., Premstaller, A., Gebhardt, C., and Oefner, P. J. (2002). Genotyping of SNPs in a polyploid genome by pyrosequencing. *Biotechniques*, 32(3):592–603.

Rijksmuseum Kröller-Müller (1961). *Catalogus bij de tentoonstelling van 272 werken van Vincent van Gogh behorende tot de verzameling van het Rijksmuseum Kröller-Müller in Otterlo [1949]*. Rijksmuseum Kröller-Müller, Otterloo, 5th edition.

Roldán, I., Wattebled, F., Mercedes Lucas, M., Delvallé, D., Planchot, V., Jiménez, S., Pérez, R., Ball, S. G., D'Hulst, C., and Mérida, A. (2007). The phenotype of soluble

starch synthase IV defective mutants of *Arabidopsis thaliana* suggests a novel function of elongation enzymes in the control of starch granule formation. *The Plant Journal*, 49(3):492–504.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365.

Scaglione, D., Acquadro, A., Portis, E., Tirone, M., Knapp, S. J., and Lanteri, S. (2012). RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *BMC Genomics*, 13:3.

Schäfer-Pregl, R., Ritter, E., Concilio, L., Hesselbach, J., Lovatti, L., Walkemeier, B., Thelen, H., Salamini, F., and 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(5-6):834–846.

Scheidig, A., Fröhlich, A., Schulze, S., Lloyd, J. R., and Kossmann, J. (2002). Downregulation of a chloroplast-targeted β-amylase leads to a starch-excess phenotype in leaves. *Plant Journal*, 30(5):581–591.

Sharma, R., Goossens, B., Kun-Rodrigues, C., Teixeira, T., Othman, N., Boone, J. Q., Jue, N. K., Obergfell, C., O'Neill, R. J., and Chikhi, L. (2012). Two different high throughput sequencing approaches identify thousands of de novo genomic markers for the genetically depleted Bornean elephant. *PLOS ONE*, 7(11):e49533.

Sharma, S. K., Bolser, D., de Boer, J., Sonderkaer, M., Amoros, W., Carboni, M. F., D'Ambrosio, J. M., de la Cruz, G., Di Genova, A., Douches, D. S., Eguiluz, M., Guo, X., Guzman, F., Hackett, C. A., Hamilton, J. P., 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, S. J., Torres, Y., Waugh, R., Zhang, Z., Huang, S., Visser, R. G. F., Bachem, C. W. B., Sagredo, B., Feingold, S. E., Orjeda, G., Veilleux, R. E., Bonierbale, M., Jacobs, J. M. E., Milbourne, D., Martin, D. M. A., and Bryan, G. J. (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.

Simko, I. (2004). One potato, two potato: haplotype association mapping in autotetraploids. *Trends in Plant Science*, 9(9):441–448.

Simko, I., Costanzo, S., Haynes, K. G., Christ, B. J., and Jones, R. W. (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.

Simko, I., Haynes, K. G., and Jones, R. W. (2006). Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. *Genetics*, 173(4):2237–45.

Singh, P., Kaloudas, D., and Raines, C. A. (2008). Expression analysis of the *Arabidopsis* CP12 gene family suggests novel roles for these proteins in roots and floral tissues. *Journal of Experimental Botany*, 59(14):3975–3985.

Skot, L., Humphreys, M. O., Armstead, I., Heywood, S., Skot, K. P., Sanderson, R., Thomas, I. D., Chorlton, K. H., and Hamilton, N. R. S. (2005). An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). *Molecular Breeding*, 15(3):233–245.

Soto-Cerda, B. J. and Cloutier, S. (2012). Association mapping in plant genomes. In Caliskan, M., editor, *Genetic Diversity in Plants*, chapter 3, pages 29–54. InTech, online.

Sowokinos, J. R., Vigdorovich, V., and Abrahamsen, M. (2004). Molecular cloning and sequence variation of UDP-glucose pyrophosphorylase cDNAs from potatoes sensitive and resistant to cold sweetening. *Journal of Plant Physiology*, 161(8):947–955.

Spooner, D. M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G. J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14694–9.

Statistisches Bundesamt (2012). *Statistisches Jahrbuch über Ernährung, Landwirtschaft und Forsten der Bundesrepublik Deutschland*.

Stich, B. and Melchinger, A. E. (2009). Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. *BMC Genomics*, 10(1):94.

Stich, B. and Melchinger, A. E. (2010). An introduction to association mapping in plants. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, 5(39):1–9.

Stich, B., Möhring, J., Piepho, H.-P., Heckenberger, M., Buckler, E. S., and Melchinger, A. E. (2008). Comparison of mixed-model approaches for association mapping. *Genetics*, 178(3):1745–54.

Stich, B., Urbany, C., Hoffmann, P., and 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(6):718–724.

Südstärke Informationsdienst (2013a). Kartoffelstärke. Eine besondere Stärke: Viskositätsverhalten.

Südstärke Informationsdienst (2013b). Stärken für den technischen Einsatz.

Südstärke Informationsdienst (2013c). Stärken für die Nahrungsmittelindustrie.

Südstärke Informationsdienst (2013d). Stärken für die Papierindustrie.

Südstärke Informationsdienst (2013e). Stärken für die Textilindustrie.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S., and Terauchi, R. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal*, 74(1):174–83.

Tauberger, E., Fernie, A. R., Emmermann, M., Renz, A., Kossmann, J., Willmitzer, L., and Trethewey, R. N. (2000). Antisense inhibition of plastidial phosphoglucomutase provides compelling evidence that potato tuber amyloplasts import carbon from the cytosol in the form of glucose-6-phosphate. *The Plant Journal*, 23(1):43–53.

The Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475:189–195.

Trick, M., Adamski, N. M., Mugford, S. G., Jiang, C.-C., Febrer, M., and Uauy, C. (2012). Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biology*, 12(1):14.

Truong, H. T., Ramos, A. M., Yalcin, F., De Ruiter, M., Van der Poel, H. J. A., Huvenaars, K. H. J., Hogers, R. C. J., Van Enckevort, L. J. G., Janssen, A., Van Orsouw, N. J., and Van Eijk, M. J. T. (2012). Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLOS ONE*, 7(5):e37565.

Uematsu, K., Suzuki, N., Iwamae, T., Inui, M., and Yukawa, H. (2012). Increased fructose 1,6-bisphosphate aldolase in plastids enhances growth and photosynthesis of tobacco plants. *Journal of Experimental Botany*, 63(8):3001–3009.

Uitdewilligen, J. (2012). *Discovery and Genotyping of Existing and Induced DNA Sequence Variation in Potato*. Phd thesis, Wageningen University.

Uitdewilligen, J. G. A. M. L., Wolters, A.-M. A., D'hoop, B. B., Borm, T. J. A., Visser, R. G. F., and Van Eck, H. J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLOS ONE*, 8(5):e62355.

Urbany, C., Colby, T., Stich, B., Schmidt, L., Schmidt, J., and Gebhardt, C. (2012). Analysis of natural variation of the potato tuber proteome reveals novel candidate genes for tuber bruising. *Journal of Proteome Research*, 11(2):703–16.

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., and Gebhardt, C. (2011). Association genetics in *Solanum tuberosum* provides new insights into potato tuber bruising and enzymatic tissue discoloration. *BMC Genomics*, 12(1):7.

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

Van den Berg, R. G. and Jacobs, M. M. (2007). Molecular taxonomy. In Vreugdenhil, D., Bradshaw, J., Gebhardt, C., Govers, F., Taylor, M. A., MacKerron, D. K., and Ross, H. A., editors, *Potato Biology and Biotechnology: Advances and Perspectives*, chapter 4, pages 55–76. Elsevier, London.

Van Eck, H. J. (1995). *Localisation of morphological traits on the genetic map of potato using RFLP and isozyme markers*. Phd thesis, Landbouwhogeschool, Wageningen University.

Van Eck, H. J. (2007). Genetics of morphological and tuber traits. In Vreugdenhil, D., Bradshaw, J., Gebhardt, C., Govers, F., Taylor, M. A., MacKerron, D. K., and Ross, H. A., editors, *Potato Biology and Biotechnology: Advances and Perspectives*, chapter 6, pages 91–115. Elsevier, London.

Van Os, H., Andrzejewski, S., Bakker, E., Barrena, I., Bryan, G. J., Caromel, B., Ghareeb, B., Isidore, E., De Jong, W., Van Koert, P., Lefebvre, V., Milbourne, D., Ritter, E., Rouppe van der Voort, J. N. A. M., Rousselle-Bourgeois, F., Van Vliet, J., Waugh, R., Visser, R. G. F., Bakker, J., and Van Eck, H. J. (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(2):1075–1087.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–23.

Von Scheele, C., Svensson, G., and Rasmusson, J. (1936). Die Bestimmung des Stärkegehalts und der Trockensubstanz der Kartoffel mit Hilfe des spezifischen Gewichts. *Landwirtschaftliche Versuchs-Stationen*, pages 67–96.

Voorrips, R. E., Gort, G., and Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*, 12(1):172.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., and Kuiper, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, 23:4407–4414.

Wang, N., Fang, L., Xin, H., Wang, L., and Li, S. (2012). Construction of a high-density genetic map for grape using next generation restriction-site associated DNA sequencing. *BMC Plant Biology*, 12(1):148.

Wang, X. Q., Zhao, L., Eaton, D. a. R., Li, D. Z., and Guo, Z. H. (2013). Identification of SNP markers for inferring phylogeny in temperate bamboos (Poaceae: Bambusoideae) using RAD sequencing. *Molecular Ecology Resources*, 13(5):938–45.

Weckx, S., Del-Favero, J., Rademakers, R., Claes, L., Cruts, M., De Jonghe, P., Van Broeckhoven, C., and De Rijk, P. (2005). novoSNP, a novel computational tool for sequence variation discovery. *Genome Research*, 15(3):436–442.

Wedel, N., Soll, J., and Paap, B. K. (1997). CP12 provides a new mode of light regulation of Calvin cycle activity in higher plants. *Proceedings of the National Academy of Sciences*, 94(September):10479–10484.

Welsh, J. and McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, 18(24):7213–8.

Werij, J., Kloosterman, B. A., Celis-Gamboa, C., de Vos, C., America, T., Visser, R. G. F., and Bachem, C. W. B. (2007). Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, QTL, and expression analysis. *Theoretical and Applied Genetics*, 115(2):245–252.

Werij, J. S., Furrer, H., Van Eck, H. J., Visser, R. G. F., and Bachem, C. W. B. (2012). A limited set of starch related genes explain several interrelated traits in potato. *Euphytica*, 186(2):501–516.

Wiegand, G. and Remington, S. J. (1986). Citrate synthase: structure, control, and mechanism. *Annual Review of Biophysics and Biophysical Chemistry*, 15:97–117.

Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., and Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research*, 18(22):6531–5.

Willing, E.-M., Hoffmann, M., Klein, J. D., Weigel, D., and Dreyer, C. (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics*, 27(16):2187–2193.

Wilson, L. M., Whitt, S. R., Ibanez, A. M., Rocheford, T. R., Goodman, M. M., and Buckler, E. S. (2004). Dissection of maize kernel composition and starch production by candidate gene association. *The Plant Cell*, 16(10):2719–2733.

Wolters, A.-M. A., Uitdewilligen, J. G. A. M. L., van Eck, H. J., De Vetten, N. C. M. H., and Visser, R. G. F. (2011). Method for modulating the level of phosphorylation of starch in a plant line, method for selecting a plant or part thereof, including a seed and tuber, and use thereof (*Patent application*).

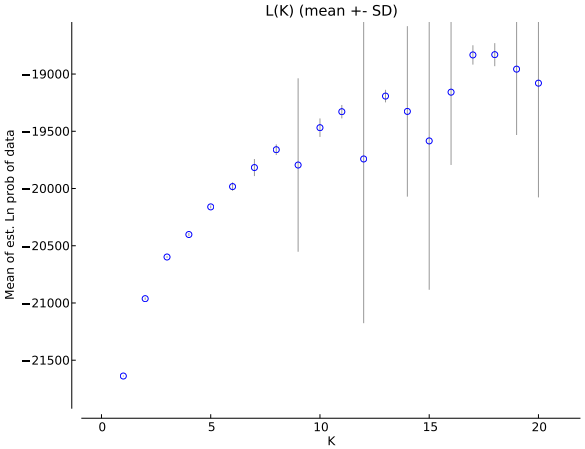
Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

Zhang, G.-L., Chen, L.-Y., Xiao, G.-Y., Xiao, Y.-H., Chen, X.-B., and Zhang, S.-T. (2009). Bulked segregant analysis to detect QTL related to heat tolerance in rice (*Oryza sativa* L.) using SSR markers. *Agricultural Sciences in China*, 8(4):482–487.

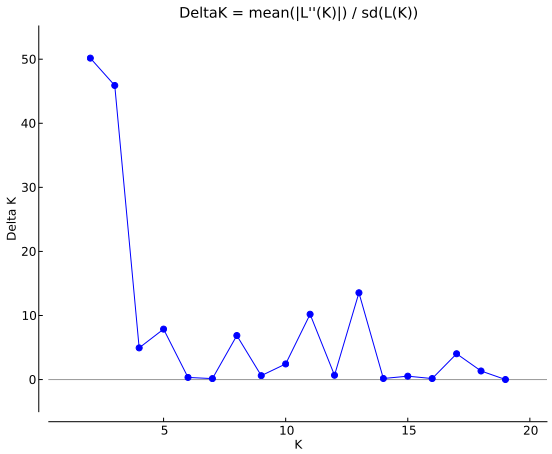
Zhang, L., Häusler, R. E., Greiten, C., Hajirezaei, M.-R., Haferkamp, I., Neuhaus, H. E., Flügge, U.-I., and Ludewig, F. (2008). Overriding the co-limiting import of carbon and energy into tuber amyloplasts increases the starch content and yield of transgenic potato plants. *Plant Biotechnology Journal*, 6(5):453–464.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42:355–360.

A. Supplemental data

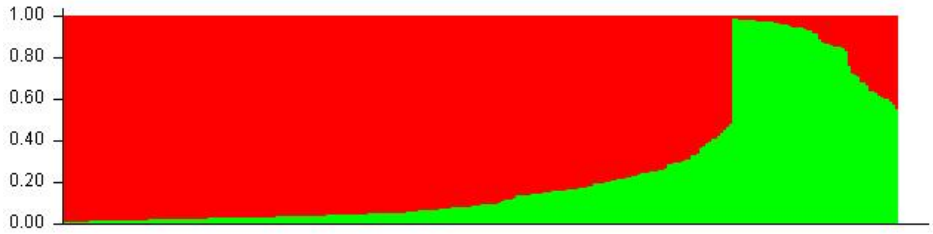


(a) Log likelihood

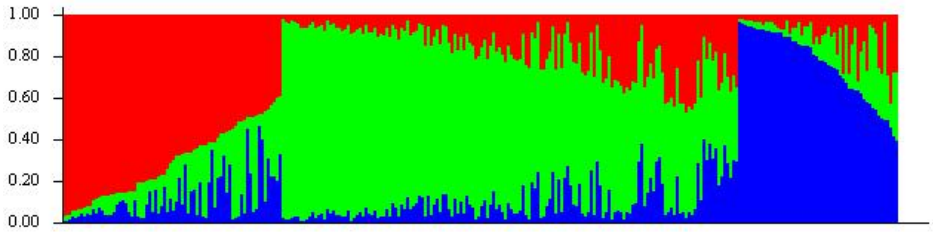


(b) Evanno method delta K

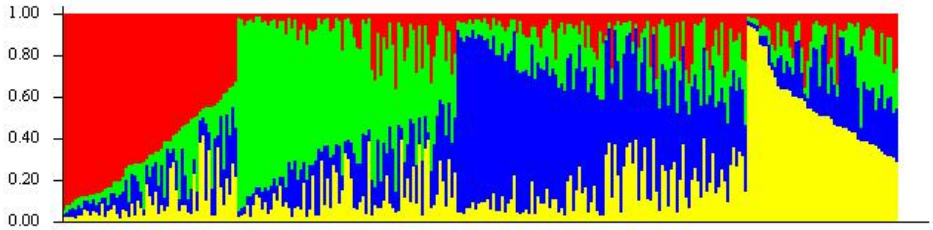
Figure A.1.: Structure harvester **output for Bayesian clustering of** Quest **population.** (a) plotted mean of log likelihood and (b) figure delta K as described by Evanno et al. (2005)



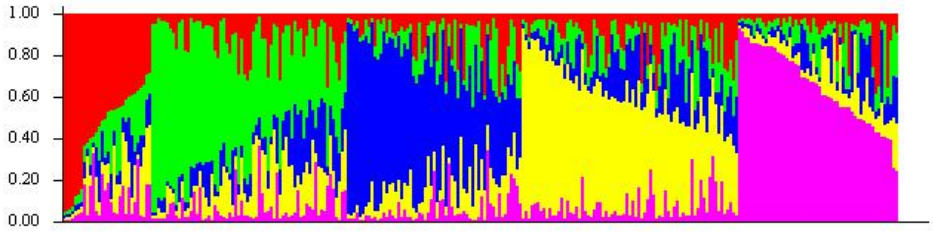
(a) Barplot of inferred subpopulations K=2 ordered according to Q-value



(b) Barplot of inferred subpopulations K=3 ordered according to Q-value



(c) Barplot of inferred subpopulations K=4 ordered according to Q-value



(d) Barplot of inferred subpopulations K=5 ordered according to Q-value

Figure A.2.: STRUCTURE **graphical output of population structure for 282 genotypes for subpopulations K=2, K=3, K=4 and K=5.** Genotypes are ordered according to the probabilities (Q-values) of each genotype belonging to one of the number of inferred subpopulations

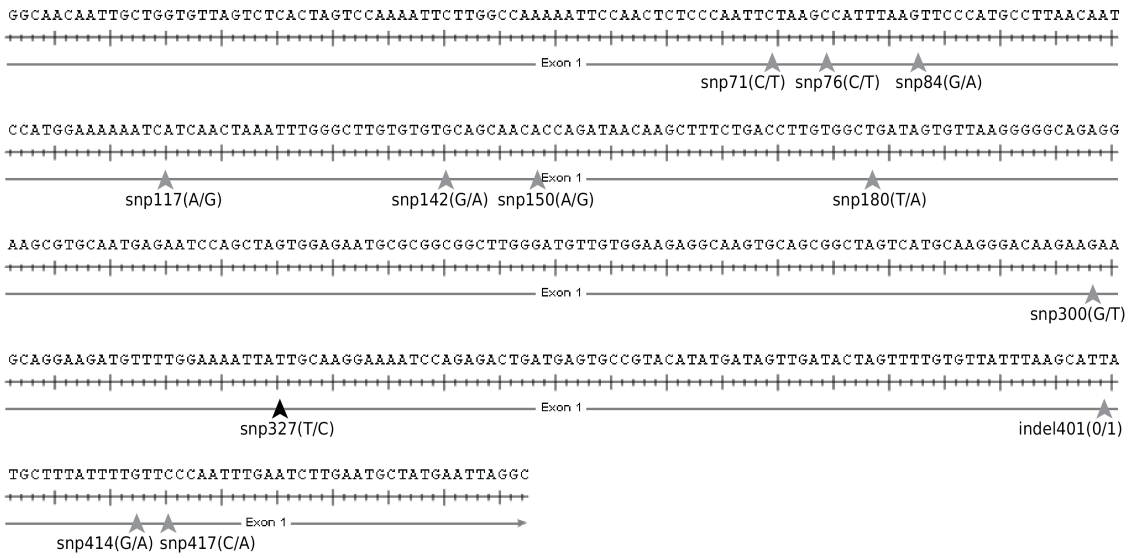


Figure A.3.: **Amplicon sequence *CP12-2*.** The genomic sequence of the PCR fragment that was analyzed for association mapping is displayed. Detected SNPs and indels are indicated by an arrow. Arrows of significantly (α=0.01) associated sequence polymorphisms are in black. The positions are given in relation to the ATG

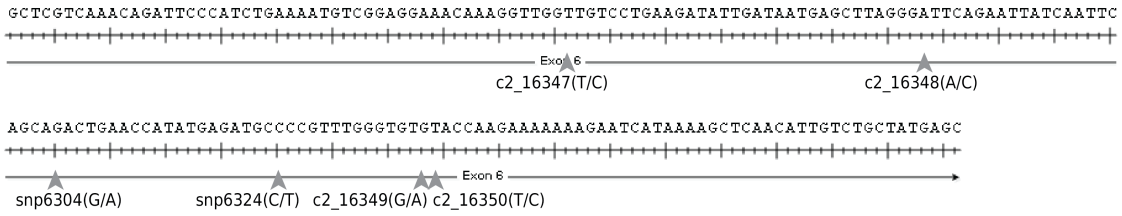


Figure A.4.: **Amplicon sequence *AP2TF*.** For detailed description see Figure A.3

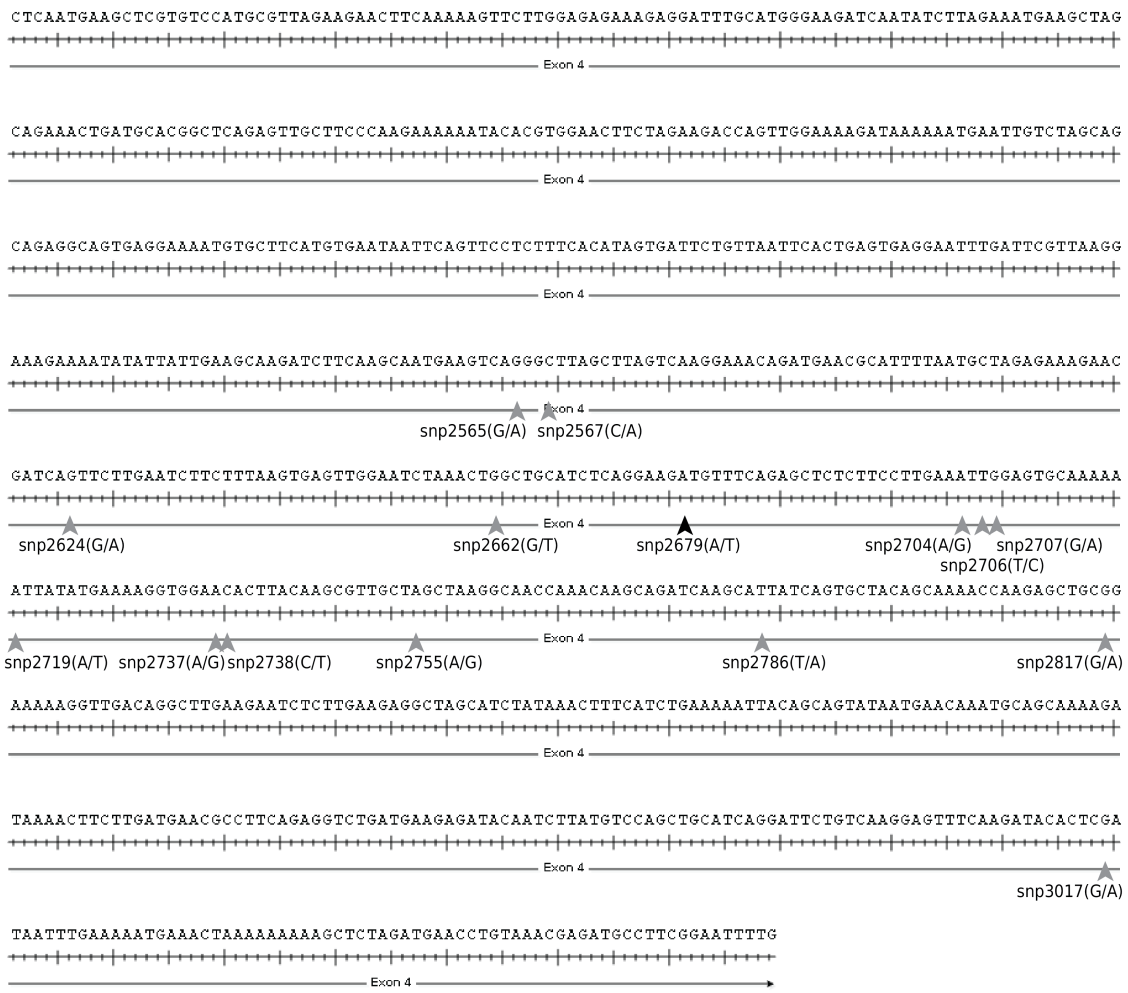


Figure A.5.: **Amplicon sequence *SSsIV*.** For detailed description see Figure A.3

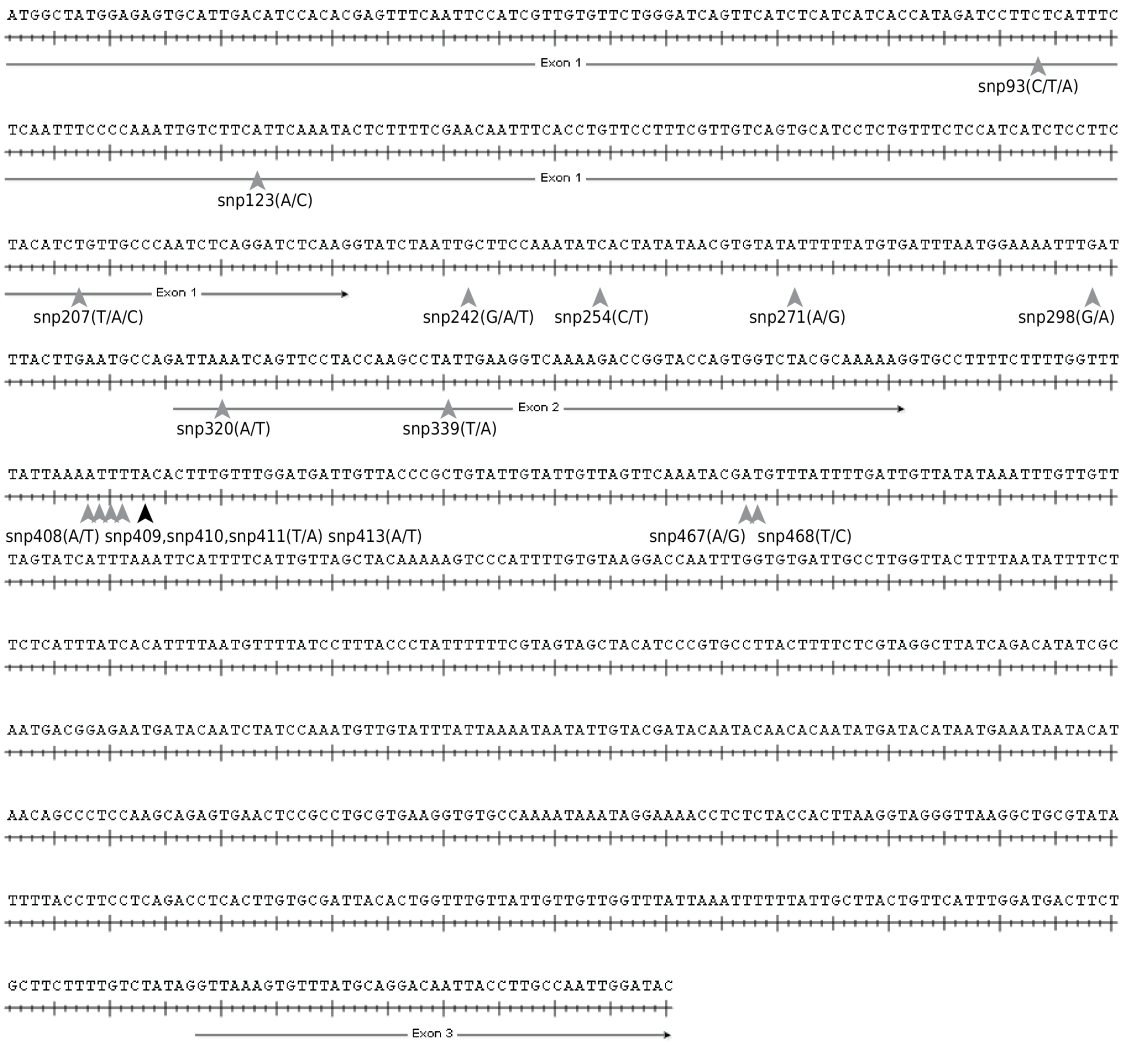


Figure A.6.: **Amplicon sequence *PGM1*.** For detailed description see Figure A.3

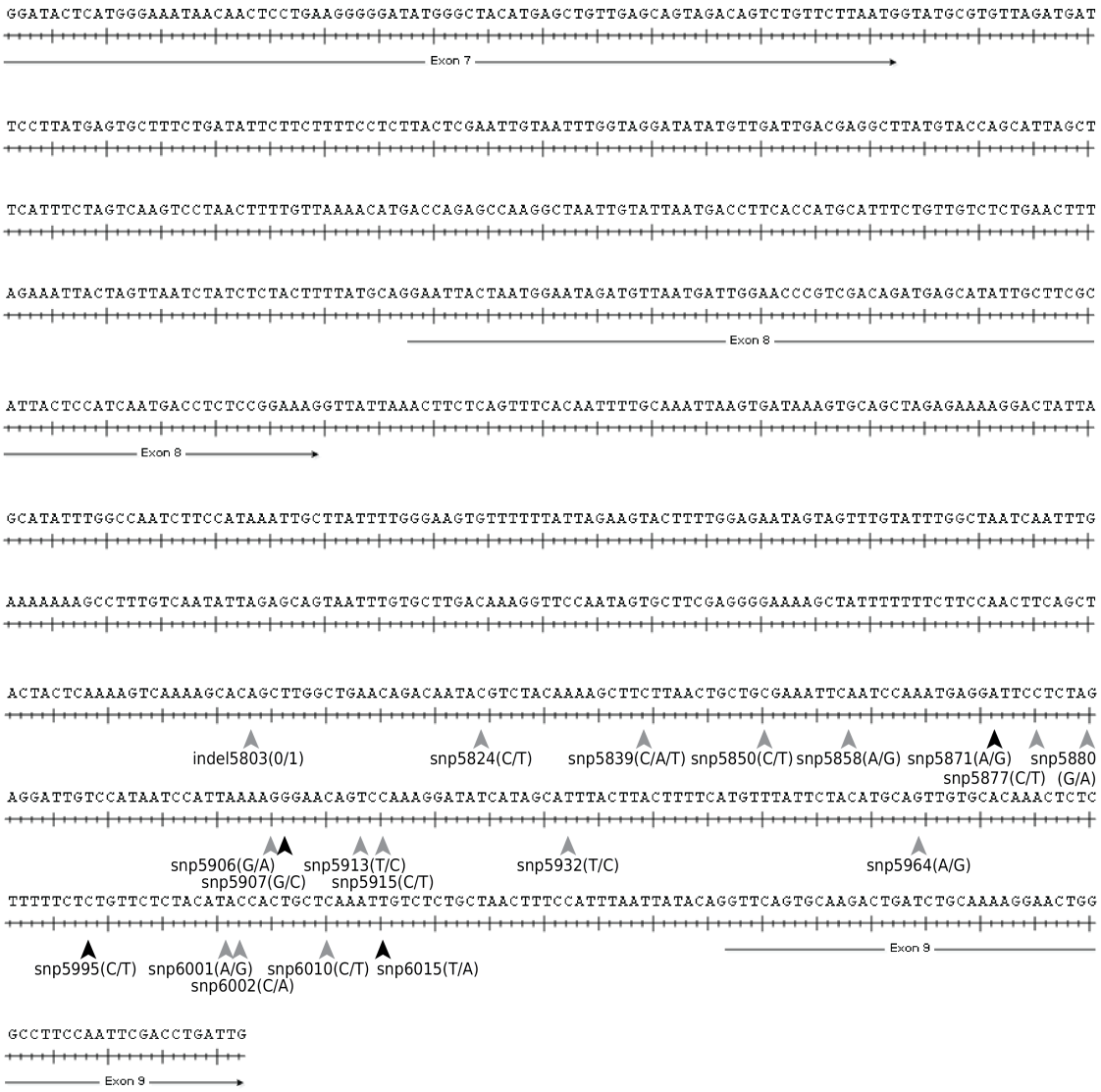


Figure A.7.: **Amplicon sequence *SssI*.** For detailed description see Figure A.3

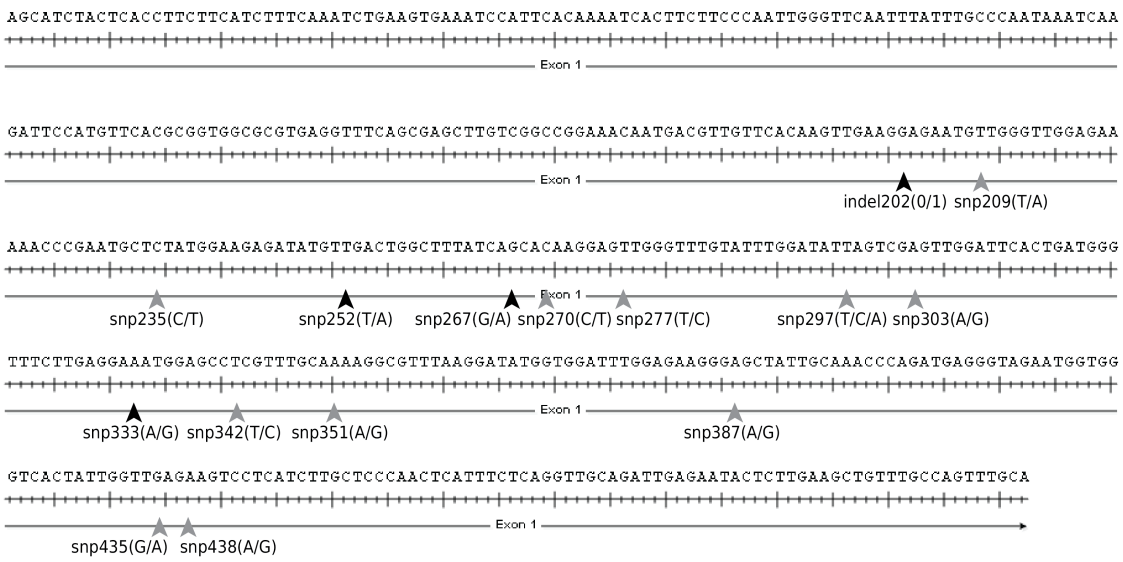


Figure A.8.: **Amplicon sequence *PGI1*.** For detailed description see Figure A.3

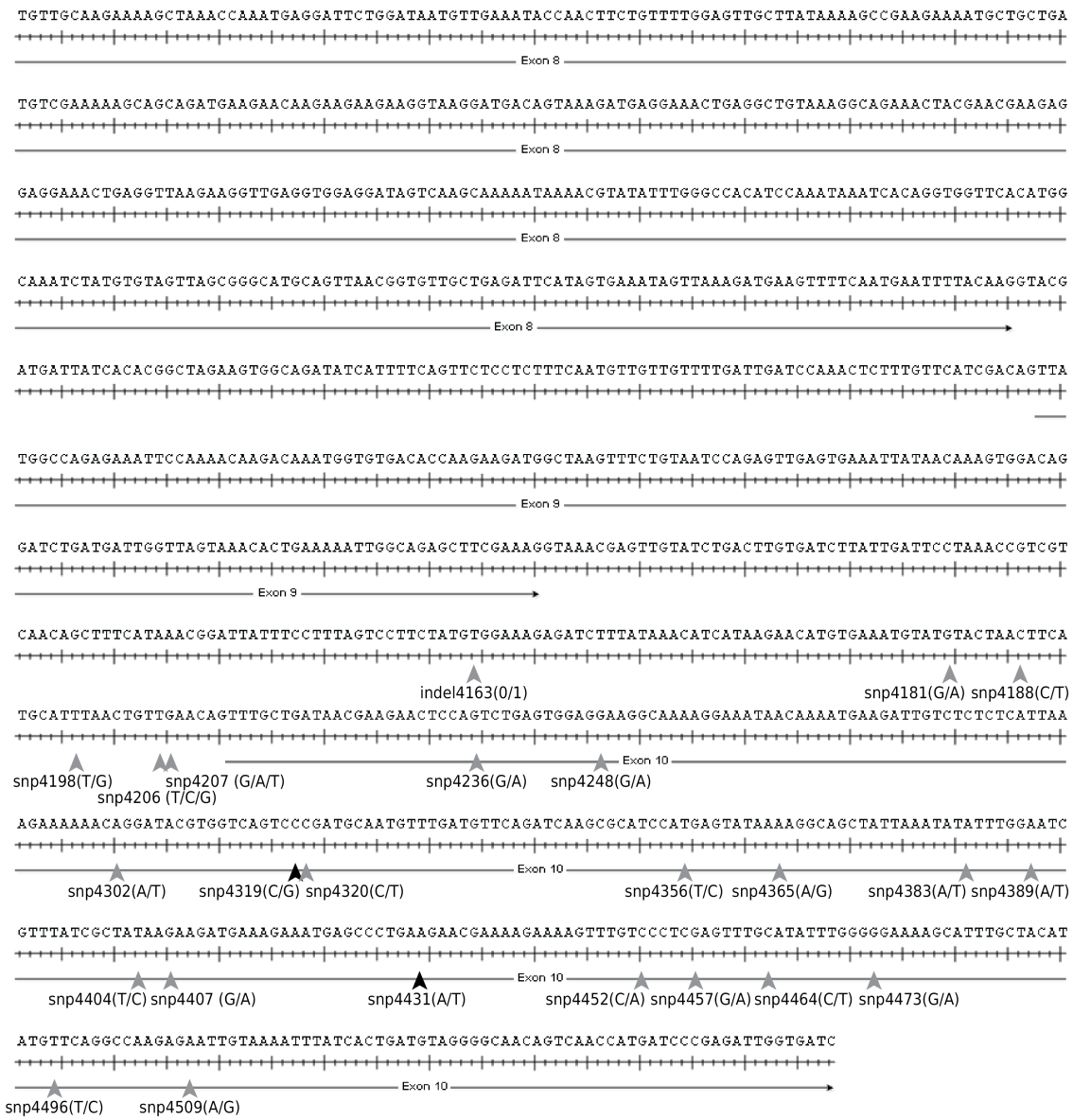


Figure A.9.: **Amplicon sequence *Pho1b*.** For detailed description see Figure A.3

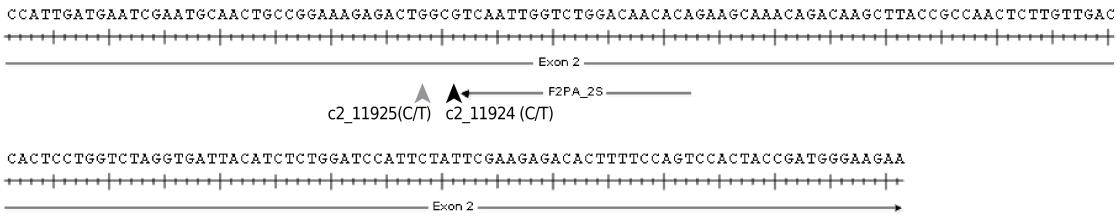


Figure A.10.: **Amplicon sequence *F2PA*.** For detailed description see Figure A.3

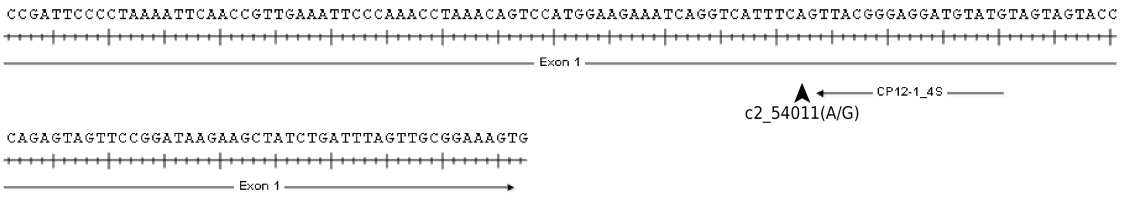


Figure A.11.: **Amplicon sequence *CP12-1*.** For detailed description see Figure A.3

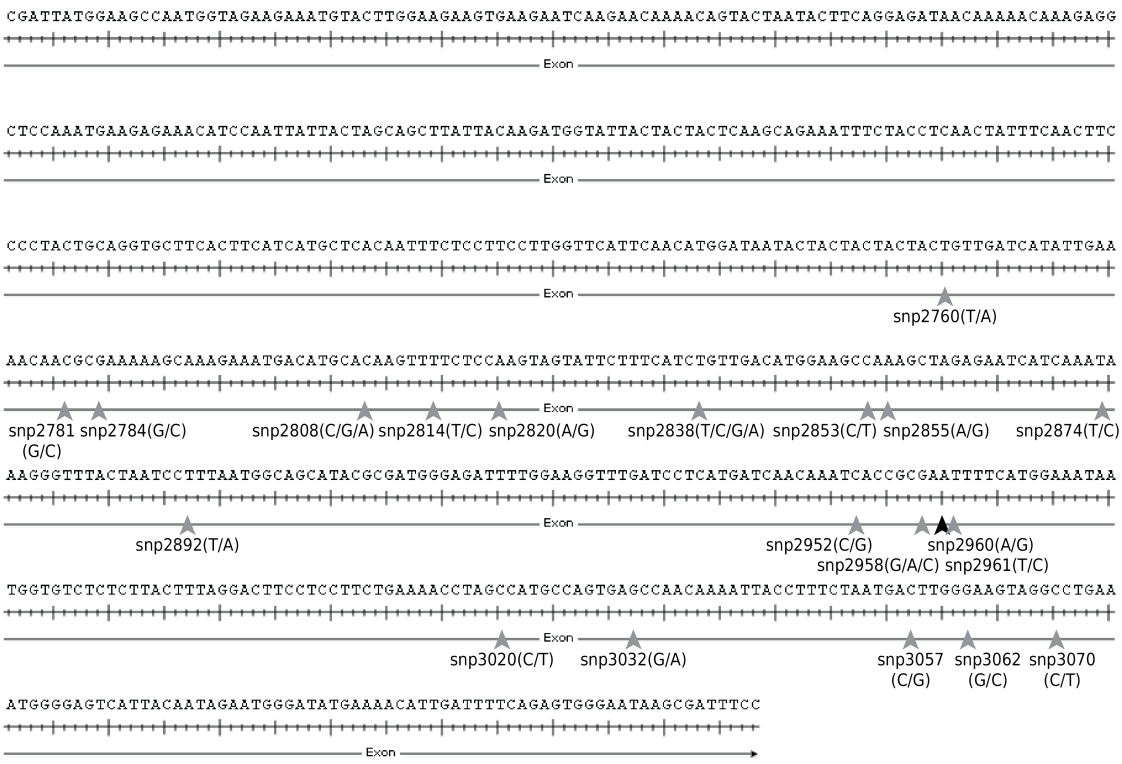


Figure A.12.: **Amplicon sequence *BEL5*.** For detailed description see Figure A.3

Figure A.13.: **Amplicon sequence *QUA1*.** For detailed description see Figure A.3

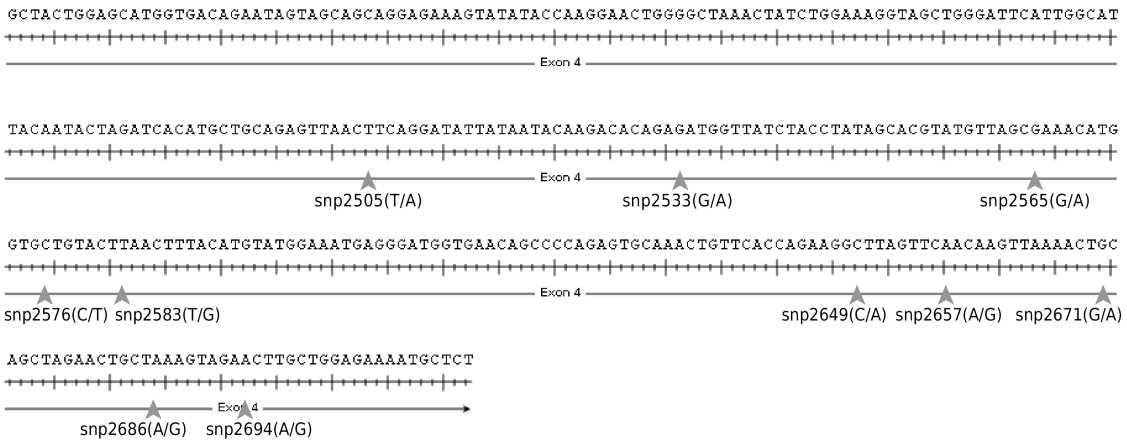


Figure A.14.: **Amplicon sequence *BMY1*.** For detailed description see Figure A.3

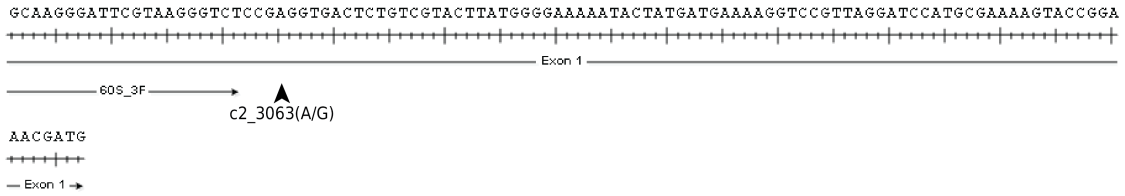


Figure A.15.: **Amplicon sequence *60S*.** For detailed description see Figure A.3

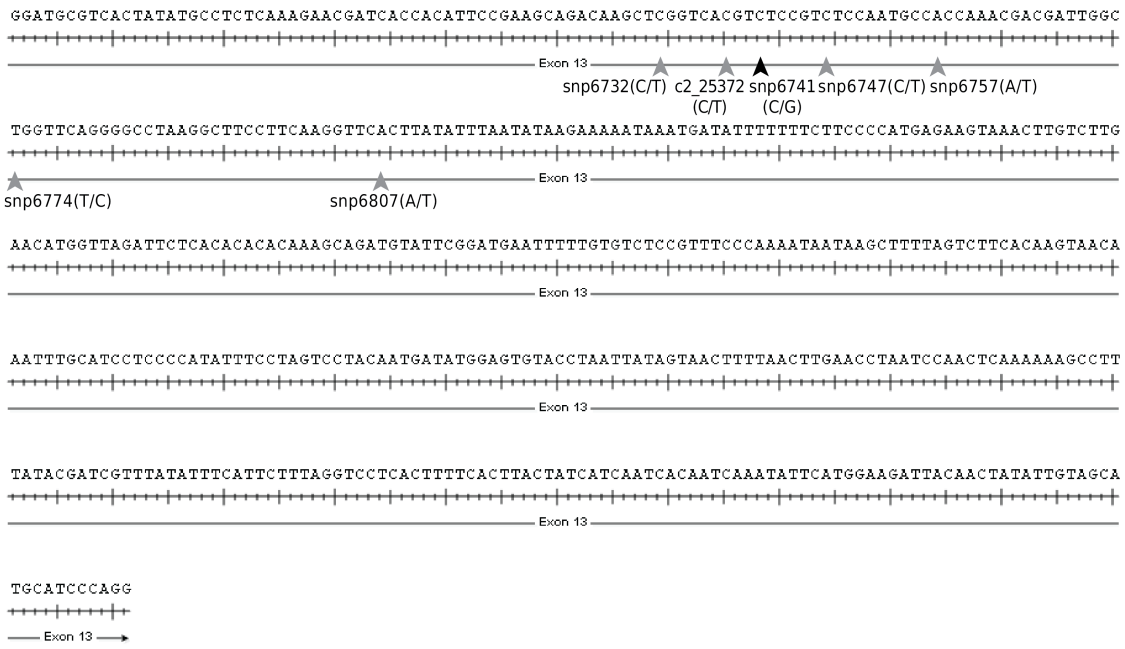


Figure A.16.: **Amplicon sequence *CIS*.** For detailed description see Figure A.3

Table A.1.: **Genotype information accompanied by their code used in the** QUEST **project**. Codes are as following. First digit: S=Standard, grown at breeding stations Appacale and Neiker, A=breeding station Appacale, N=breeding station Neiker; second digit: 0+1=commercial variety, 2+3=breeding clone, 4=landrace variety, 5=diploid clone, 6=Wild *Solanum* species; third digit: sequential number

Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone
S001	Agria	A039	Sahel	A520	D99H3.2	N099	Voyager
S002	Alava	A040	Sandra	N001	Adriana	N100	Zadorra
S003	Alegria Oro	A041	Shannon	N002	Agata	N101	Zafira
S004	Almera	A042	Surya	N003	Aladin	N103	Zarina
S005	Amanda	A043	Tosca	N004	Albata	N104	Zela
S006	Antina	A044	Tuskar	N005	Alda	N105	Zepa
S007	Baraka	A045	Ute	N006	Amalia	N106	Zunta
S008	Camila	A046	Valeria	N007	Ambition	N201	05/104-4
S009	Candela	A047	VR 808	N008	Amora	N202	05/110-4
S010	Carrera	A201	2006P46-6	N009	Amorosa	N203	05/111-2
S011	Daniela	A202	05/104-2	N010	Andean Sunrise	N204	05/112-3
S012	Desiree	A203	05/104-5	N011	Arcade	N205	05/113-2
S013	Esta	A204	05/110-3	N012	Arene	N206	05/114-3
S014	Fina de Carvallo	A205	05/112-1	N013	Armada	N207	05/115-4
S015	Fina de Gredos	A206	05/112-2	N014	Arrow	N208	05/116-3
S016	Fontane	A207	05/113-5	N015	Artemis	N209	05/117-1
S017	Goya	A208	05/114-2	N016	Asun	N210	05/118-1
S018	H-88 31/34	A209	05/115-1	N017	Avalon	N211	05/119-2
S019	H98A/11	A210	05/116-4	N018	Ayala	N212	05/122-3
S020	H98A/18	A211	05/117-2	N019	Belda	N213	05/123-1

Table A.1.: (continued)

Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone
S021	H98A/25	A212	05/117-5	N020	Brda	N401	Senora Warni
S022	H98B/2	A213	05/118-3	N021	Buesa	N402	Chaucha
S023	Heidrun	A214	05/119-3	N022	Cabadia	N403	Socco Huaccoto
S024	Idoia	A215	05/119-5	N024	Cazona	N404	Sipancachi
S025	Isle of Jure	A216	05/122-1	N025	Cherie	N405	Laram Ajawiri
S026	Jaerla	A217	05/122-5	N026	Corine	N406	NKD-141
S027	Jimena	A218	05SH74-4	N027	Corrida	N407	Yana Sucre
S028	Kennebec	A219	05SH83-7	N028	Diba	N408	Wila Huaka Lajra
S029	Lady Claire	A220	2003P55-6	N029	Duquesa	N409	Rosca
S030	Leire	A221	2000P4-6	N030	Edurne	N410	Chiar Surimana O Phinu
S031	Melibea	A222	2000P5-5	N031	Elfe	N411	Morada Turuna
S032	Miranda	A223	2000Q84-7	N032	Europrima	N412	Kashpadana Amarilla
S033	Monalisa	A224	2001Q29-10	N033	Fenix	N413	Holandesa
S034	Mustang	A225	2003P32-4	N034	Festival	N414	Poluya
S035	Nagore	A226	2003P54-4	N035	Gorbea	N415	Negrita
S036	Nela	A227	2004M30-5	N036	Ibicenca	N416	Color Unckuna
S037	Opal	A228	2004P10-10	N040	Inona	N417	NKD-168
S038	Orla	A229	2004P13-6	N041	Integra	N418	NKD-169
S039	Palogan	A230	2004P8-4	N042	Irati	N601	AMB 1643
S040	Panda	A231	2004Q28-7	N043	Isla	N602	ADR 18344
S041	Priamos	A232	2004Q3-3	N044	Iturrieta	N603	BER 3607
S042	Ramses	A233	2004R81-1	N045	Jesus	N604	BLV 2190
S043	Spinta	A234	2005P107-3	N046	Kasta	N605	BST 7986

Table A.1.: (continued)

Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone
S044	Stemster	A235	2005P135-2	N047	Kondor	N606	BRC 28023
S045	Turia	A236	2005P144-1	N048	L 37 (4x)	N607	BRD 218215
S046	Valetta	A237	2005P88-5	N049	Lora	N608	BUK 15424
S047	Valnera	A238	2005P89-4	N050	LT-8	N609	BLB 8008
S048	Verdi	A239	2005Q102-2	N051	LT-9	N610	CAN 501
S049	Vivaldi	A240	2005Q106-3	N052	Lutetia	N611	CAP 32678
S050	Zorba	A241	2005Q49-3	N053	Madeleine	N612	CPH 10052
A001	Anais	A242	2005Q69-2	N054	Maika	N613	CMM 5856
A002	Aster	A243	2006D9-6	N055	Maniton	N614	CHC 17034
A003	Asterix	A244	2006FF67-17	N056	Marfona	N615	CHP 18060
A004	Atlantic	A245	2006P37-2	N057	Marietema	N616	DDS 2880
A005	Aurea	A246	2006P39-11	N058	Melody	N617	FLH 2335
A006	Barna	A247	2006P42-13	N059	Merida	N618	GND 2423
A007	Bartina	A248	2006P45-2	N062	Montico	N619	GRL 7180
A008	Bellini	A249	2006P56-3	N063	Morada	N620	HAN 2843
A009	Brodick	A250	2006Q113-12	N064	Murato	N621	HDM 2848
A010	Caesar	A251	2006Q49-6	N065	Musica	N622	JAM 481
A011	Cara	A252	2006Q66-9	N066	N-180	N623	LPH 27215
A012	Chopin	A253	2006Q68-4	N067	Naga	N624	MAG 2118
A013	Cinja	A254	2006Q88-2	N068	Nerea	N625	MRN 2278
A014	Colorado	A255	2006Q88-25	N070	Omega	N626	MED 2226
A015	Courage	A256	2006Q88-34	N071	Onda	N627	MCD 17596
A016	Denar	A257	2006Q89-1	N072	Orchestra	N628	MCQ 2319

Table A.1.: (continued)

Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone	Code	Variety/Clone
A017	Ditta	A258	2006Q92-2	N073	Pecaro	N629	MTP 17829
A018	Electra	A259	99P41-6	N074	Pedro Muñoz	N630	NRS 7211
A019	Fabiola	A260	SH 2267	N075	Presto	N631	OKA 17998
A020	Fabula	A501	03-DH2-1/26	N076	Primavera	N632	PTA 15442
A021	Gabriella	A502	03-DH2-1/29	N077	Red Baron	N633	PCS 2877
A022	Granola	A503	05-DH2-1/48	N078	Red Pontiac	N634	PHU IPV48
A023	Hermes	A504	04-DGNR13-7	N079	Riviera	N635	PNT 3863
A024	Husar	A505	04-DGNR13-9	N080	Roja Riñon	N636	PLD 8182
A025	Innovator	A506	04-DGNR14-2	N081	Romano	N637	PLT 53650
A026	Labadia	A507	04-DGNR14-3	N083	Rosa Gold	N638	QUM 27163
A027	Lady Rosetta	A508	04-DGNR15-4	N084	Rudolph	N639	RAP 636
A028	Laura	A509	04-DGNR15-6	N085	Saline	N640	SPL 959
A029	Marella	A510	04-DGNR16-4	N086	San	N641	STN 4715
A030	Norchip	A511	04-DIP1/1	N087	Satellite	N642	STO 2092
A031	Obelix	A512	04-DIP1/3	N089	Saviola	N643	SCR 18206
A032	Optima	A513	DIP1/9	N091	Sofia	N644	TAR 3746
A033	Pamela	A514	05-DGNR17-1	N092	Soprano	N645	TOR 767
A034	Ponto	A515	05-DH2-12-2	N094	Taurus	N646	TRF 18335
A035	Ramos	A516	05-DH2-14-3	N095	Tebina	N647	VNT 8239
A036	Red Scarlett	A517	05-DH2-17-1	N096	Tramontana	N648	VRN 15451
A037	Rodeo	A518	05-DH2-17-2	N097	Victor	N649	VER 1340
A038	Romeo	A519	D99H2.10	N098	Violeta	N650	YUN 2173

B. Supplemental data provided on CD

Supplemental	Content
Figure B.1	Electrophoresis patterns of SSR markers in QUEST
Protocol B.1	Pyrosequencing protocol (version august 2013)
Protocol B.2	RAD sequencing protocol (version august 2012)
Sequence B.1	Genomic sequence *CP12-2* (PGSC0003DMG400009042)
Sequence B.2	Genomic sequence *AP2TF* (not annotated)
Sequence B.3	Genomic sequence *SSsIV* (PGSC0003DMG400008322)
Sequence B.4	Genomic sequence *PGM1* (not annotated)
Sequence B.5	Genomic sequence *SssI* (PGSC0003DMG402018552)
Sequence B.6	Genomic sequence *PGI1* (PGSC0003DMG400012910)
Sequence B.7	Genomic sequence *Pho1b* (PGSC0002DMG400028382)
Sequence B.8	Genomic sequence *F2PA* (PGSC0003DMG400030565)
Sequence B.9	Genomic sequence *CP12-1* (PGSC0003DMG400007286)
Sequence B.10	Genomic sequence *BEL5* (PGSC0003DMG400005930)
Sequence B.11	Genomic sequence *QUA1* (PGSC0003DMG400020103)
Sequence B.12	Genomic sequence *BMY1* (PGSC0003DMG400001855)
Sequence B.13	Genomic sequence *60S* (PGSC0003DMG400029622)
Sequence B.14	Genomic sequence *CIS* (PGSC0003DMG400007797)
Table B.1	List of genotypes that were grown in the QUEST project
Table B.2	Primary phenotype data
Table B.3	Adjusted entry means QUEST population and diploid clones
Table B.4	Partial correlations between phenotypic traits in the QUEST population
Table B.5	SSR markers genotyped in the QUEST population
Table B.6	PCR markers genotyped in the QUEST population
Table B.7	Genotypic data QUEST population (Chapters 2+ 3)
Table B.8	All marker-trait associations (p-value <0.05)
Table B.9	Comparison between GLM, MLM-k and MLM-Pk statistical models (p-values)
Table B.10	LD between alleles of candidate gene SNPs (Chapter 2)

Table B.11	LD between alleles of candidate gene SNPs (Chapter 3)
Table B.12	LD between alleles of SolCAP SNP loci (Chapter 3)
Table B.13	Correlation between SolCAP technical replicates
Table B.14	Result chi-square test SolCAP SNPs in case-control study for TSC
Table B.15	Result chi-square test SolCAP SNPs in case-control study for TY
Table B.16	Result chi-square test SolCAP SNPs in case-control study for TSY
Table B.17	List of 328 significant (p-value <0.01) SolCAP SNPs in the case-control studies
Table B.18	Compiled list of 142 annotated candidate genes and markers from previous studies
Table B.19	List of detected RADseq SNPs in the case-control study for TSC
Table B.20	List of detected RADseq SNPs in the case-control study for TY
Table B.21	List of detected RADseq SNPs in the case-control study for TSY
Table B.22	List of significant (FDR <0.05) RADseq SNPs in annotated loci in the case-control study for TSC
Table B.23	List of significant (FDR <0.05) RADseq SNPs in annotated loci in the case-control study for TY
Table B.24	List of significant (FDR <0.05) RADseq SNPs in annotated loci in the case-control study for TSY
Table B.25	List of significant (FDR <0.05) RADseq SNPs in the TSC case-control study containing more than 5 significant and non-synonymous RADseq SNPs per locus
Table B.26	List of significant (FDR <0.05) RADseq SNPs in the TY case-control study containing more than 5 significant and non-synonymous RADseq SNPs per locus
Table B.27	List of significant (FDR <0.05) RADseq SNPs in the TSY case-control study containing more than 5 significant and non-synonymous RADseq SNPs per locus
Table B.28	List of novel candidate loci from RADseq for tuber starch content, tuber yield and starch yield, more than 5 significant and non-synonymous SNPs per locus

C. Partners involved in different parts of the work

Felisa Ortega and Claudia Lopez-Vizcon

Eckhard Tacke and Stefanie Hartje

Jost Muth

Enrique Ritter, Leire Barandalla and Ana Aragones

Christiane Gebhardt, Birgit Walkemeier, Jia Ding and Maarten Koornneef.
Master students involved in this work were Jo-Chien Liao (Wageningen University), Vicky Tilmes (Universiät Köln) and Shyamkumar Immadi (Universität Bonn)

This work was supported by the EU program PLANT-KBBE (Knowledge Based Bio-Economy) Initiative of 7th PM, under contract EUI-2009-04028.

D. Acknowledgements

At this point I thank everyone, who directly and indirectly contributed my dissertation project. I am very grateful to look back on the large number of persons that were involved scientifically as well as emotionally .
My sincere gratitude to PD Dr. Christiane Gebhardt for providing me with this interesting project. Thanks a lot for your continuous support and the freedom to make this project my own *Quest*. Also thanks to Prof. Dr. Martin Hülskamp for taking the time to be my second thesis reviewer and Prof. Dr. Ulf-Ingo Flügge for chairing the examination board. Dr. Matthias Fischer and Dr. Susanne Rossmann critically read my manuscript. Your comments and suggestions were of great help and I highly appreciate your time and input. Birgit Walkemeier provided me with essential support in the lab. Your warm, interested spirit often made my day. Thanks also to Fabio Cericola, Jo-Chien Liao, Dr. Rena Sanetomo and Vicky Tilmes for practical help in the lab and Dr. Jia Ding for the analysis of the RAD sequences. Anja Bus was a great help with setting up the RAD sequencing experiment. Thanks also to Prof. Dr. Benjamin Stich for the statistical support and Dr. Alexander Lipka for the potato GAPIT script.
My thesis project was part of the QUEST project that was funded by the BMBF. Special thanks to the QUEST collaboration partners for support, friendly ears and helpful project discussions, especially Dr. Eckhard Tacke, Dr. Jost Muth, Dr. Enrique Ritter and Dr. Felisa Ortega.
With the help of Dr. Jakob Mahl (Südstärke GmbH) I learned more about the extraction and applications of potato starch.
Many thanks to the members of the Lab Gebhardt (Meki, Lena, Camila, Astrid, Matthias, Claude, Janne, Markus, Li and Sandra) and Anja, Andreas, Frederike, Jonas, Niklas, Susanne for scientific discussions, support and the pleasant working environment.
I highly appreciate the practical support by my *thesis-support team* (Gitta, Frauke, Susanne, Annina, Hanneke, Julia, Hauskreis), especially in the final writing phase. Meinen Eltern gilt besonderer Dank. Ihr habt mich meinen eigenen Weg finden und gehen lassen. My husband, Hartmut. You have been an invaluable support and motivator on this exciting journey from the initial idea to the final manuscript. Thank you so much.
Last, but not least I am grateful to my god, who equipped me with anything I needed to complete this thesis-*Quest*.

E. Eidesstattliche Erklärung

Die vorliegende Arbeit wurde am Max-Planck-Institut für Pflanzenzüchtungsforschung in Köln durchgeführt.

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde.

Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von PD Dr. Christiane Gebhardt betreut worden.

Köln, den 11.07.2014 Elske Maria Schönhals

F. Lebenslauf

Angaben zur Person

Name	Elske Maria Schönhals
Geburtstag und -ort	27.07.1979, Kirchheimbolanden

Ausbildung

seit 06/2010	Doktorandin am Max-Planck-Institut für Pflanzenzüchtungsforschung in Köln, Arbeitsgruppe Dr. Christiane Gebhardt
10/03 - 03/09	Wageningen University (Niederlande), MSc Plant Sciences, Spezialisierung Plant Breeding and Genetic Resources
11/04 - 05/05	Charles Sturt University, Wagga Wagga (Australien)
09/01 - 02/06	Hochschule Geisenheim, Dipl.-Ing. (FH) Weinbau und Getränketechnologie, Spezialisierung Weinbau und Oenologie
2000	Abitur, Integrierte Gesamtschule Mainz-Bretzenheim, Mainz
Stipendium	Hygens Scholarship for Excellent Students 2009/2010

Berufserfahrung und Praktika

04/06 - 10/08	Versuchsingenieurin, Hochschule Geisenheim, Institut für Rebenzüchtung und Rebenveredlung, Beschreibung von Klonendiversität traditioneller deutscher Rebsorten, Selektion und Sicherung von gefährdeten genetischen Ressourcen
09/05 - 01/06	Praktikum Domaine St. Andre, Meze (Frankreich)
08/04 - 11/04	Praktikum Kölner Verbund Brauereien, Köln
02/03 - 03/03	Praktikum National Wine and Grape Industry Centre, Wagga Wagga (Australien)

Lehre

01/09 - 05/10	Tutorin für *Basic Statistics*, Wageningen University
04/06 - 10/08	Praktikum *Rebsortenkunde*, Hochschule Geisenheim
04/06 - 10/08	Praktikum *Rebenveredelung*, Berufsschule Rheingau