Discovery and Genotyping of Existing and Induced DNA Sequence Variation in Potato

Jan Uitdewilligen

Thesis committee

Thesis supervisor
Prof. dr. R.G.F. Visser
Professor of Plant Breeding
Wageningen University

Thesis co-supervisors
Dr. ir. H.J. van Eck
Assistant professor, Laboratory of Plant Breeding
Wageningen University

Dr. ir. A.M.A. Wolters
Scientist, Laboratory of Plant Breeding
Wageningen University

Other members
Prof. dr. F.A. van Eeuwijk, Wageningen University
Prof. dr. M. Groenen, Wageningen University
Prof. dr. M.E. Schranz, Wageningen University
Dr. N.C. de Vetten, Averis Seeds, Valthermond

This research was conducted under the auspices of the
Graduate School of Experimental Plant Sciences

Discovery and Genotyping of Existing and Induced DNA Sequence Variation in Potato

Jan Uitdewilligen

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 18 September 2012
at 4 p.m. in the Aula.

Jan G.A.M.L. Uitdewilligen
Discovery and genotyping of existing and induced DNA sequence variation in potato,
166 pages.

Thesis, Wageningen University, Wageningen, NL (2012)
With references, with summaries in Dutch and English

ISBN 978-94-6173-233-0

TABLE OF CONTENTS

CHAPTER 1 – General Introduction 7

CHAPTER 2 – Identification of Alleles of Carotenoid Pathway Genes Important for Zeaxanthin Accumulation in Potato Tubers 19

CHAPTER 3 – Sequence Characterization of *StGWD* Haplotypes and the Genetics of Starch Phosphate Content in Tetraploid Potato 41

CHAPTER 4 – A Next-Generation Sequencing Method for Genotyping-By-Sequencing of Highly Heterozygous Autotetraploid Potato 63

CHAPTER 5 – EMS-Induced Mutation Discovery in the M_1 Generation of Potato as a Strategy for Reverse Genetics 101

CHAPTER 6 – General Discussion 123

REFERENCES 141

SUMMARY 155

SAMENVATTING 159

DANKWOORD 163

OVER DE AUTEUR 165

CHAPTER 1

General Introduction

POTATO BREEDING

Potato *(Solanum tuberosum L.)* is a nutritious food crop, contributing carbohydrates, important amino acids and vitamins to the human diet. The species is highly heterozygous, autotetraploid (2n=4x=48), and largely reproduces vegetatively. The early adaptation of cultivated potato in Europe and North America from landraces of South American *Solanum tuberosum 'Andigenum group'* and '*Chilotanum Group'*, has been accompanied by strong artificial selection for adaptation to long days (AMES and SPOONER 2008). In the 1840s, the late blight (*Phytophthora infestans*) disease caused the "Irish potato famine", leading to subsequent mass selection for late blight resistance. This was coupled with continued selection for desirable agronomic characteristics such as vigor and yield and ultimately led to introgression of many disease resistance genes from wild species into the gene pool of European and North American potato cultivars. Current potato breeding activities are aimed at developing new high-yielding cultivars with pest and disease resistance while also meeting the increasing quality requirements of consumers and the potato processing industry. Most of the characteristics targeted for improvement are quantitative traits for which phenotypic variation depends on both genetic and environmental factors. The molecular basis of many of these traits is, however, poorly understood (HAMILTON *et al.* 2011).

MOLECULAR MARKERS

Identification and genotyping of DNA variants lies at the core of modern genetics and allows the exploration of important questions in population genetics, evolution and plant breeding (DAVEY *et al.* 2011). Molecular markers are a common tool for the development of saturated genetic and physical maps, genetic diversity analysis, genotype identification, gene isolation and the identification of loci controlling phenotypic traits, marker-assisted selection (MAS), and, more recently, genomic selection (GS). Two main types of molecular markers are fragment-based markers and the more recently developed sequence-based markers.

Traditional fragment-based molecular markers

Multiple approaches have been developed for the characterization of DNA variation in plants and to identify sequence polymorphisms using fragments of DNA. Until recently, amplified fragment length polymorphisms (AFLPs) or simple sequence repeats (SSRs) were the molecular markers of choice for DNA fingerprinting of plant genomes (ALLENDER and KING 2010; CASTILLO *et al.* 2010; D'HOOP *et al.* 2008). Both methods rely on gel-based identification of DNA fragment sizes after PCR amplification. These marker types have a number of inherent weaknesses. In AFLP and SSR-based genetic studies, all genotypes with the same SSR pattern or with the same AFLP fragment are usually considered to be identical for this marker locus. However, the internal DNA sequence composition of the AFLP or SSR marker fragment with a specific gel mobility and/or the chromosomal regions linked with a given marker fragment may exhibit significant sequence variation (GORT and VAN EEUWIJK 2012). In such cases the markers are identical by state (IBS) without being identical by descent (IBD). Furthermore, in marker-trait associations analysis, the value of a marker fragment depends on the strength of its linkage with the functional alleles of a target locus influencing phenotypic variation in the

target trait. Since SSR and AFLP markers are usually derived from non-functional, intergenic sequences, and since these markers occur at a relatively low frequency, their strong linkage with causal alleles is less likely.

Sequence-based markers

The underlying DNA sequence variants that define AFLP and SSR polymorphisms are sequence variants like single nucleotide polymorphisms (SNPs), multi-nucleotide polymorphisms (MNPs), and insertions and deletions (indels). These sequence variants are the basic units of genetic diversity and have additional value, relative to that of SSR and AFLP markers, in their ability to quantify genetic diversity and explain phenotypic variation. By generating distinct alleles, individual sequence variants, or multiple variants combined into distinct haplotypes, can be responsible for variation in specific traits or phenotypes. Sequence-based markers may also represent neutral variation that is useful for evaluating genetic diversity. Compared to multi-allelic SSR markers, an individual base change detected as a SNP is more likely to have occurred only once in evolutionary time. Owing to their greater utility in making evolutionary inferences, abundance, ease of high-throughput detection, and efficient automated genotyping, SNP markers are increasingly applied to study genetic diversity and phenotypic variation in plant breeding and basic genetics research (GRATTAPAGLIA *et al.* 2011; GRIFFIN *et al.* 2011; HUANG *et al.* 2010; LIJAVETZKY *et al.* 2007; RICKERT *et al.* 2003).

SNPs as haplotype markers

In practice, SNPs are mostly bi-allelic, even though in principle any of the four nucleotides can be present at any position in a stretch of DNA. This bi-allelic nature is due to the low frequency of mutations that lead to new SNPs. As a result, the chance for a second independent mutation that introduces a third allele at the same base position is low. Unlike for multi-allelic markers, diversity values (expected heterozygosity) of SNPs are therefore generally low. The relatively poor information content of individual SNP markers can, however, be enhanced by using haplotype markers rather than single marker scores. Haplotypes can be defined as common sets of markers in linkage phase at adjacent loci. These linked markers are further structured into haplotype blocks likely to be transmitted as a unit from generation to generation, assuming no recombination occurs (i.e., IBD). Each haplotype block contains only a few haplotypes. The minimal informative subset of SNPs associated with the haplotypes in a block are often referred to as the "haplotype tagging SNPs" or tag SNPs (GABRIEL *et al.* 2002; JOHNSON *et al.* 2001). Depending on the amount of sequence diversity, tag SNPs can be either pairwise-defined or multi-marker-defined, e.g. either a single tag SNP or a combination of tag SNPs identifies a single haplotype (DE BAKKER *et al.* 2005). Used as tag SNPs, bi-allelic SNP markers can be as informative as multi-allelic molecular markers.

When inbred plant lines are available, haplotypes can immediately be inferred from the genotype, whereas heterozygous genotypes require phase detection among the markers for accurate haplotype determination (BUNTJER *et al.* 2005). Most primary genotyping data, like SNP assay and sequencing data, will not distinguish between a pair of markers associated on

the same chromosome or located on two different homologous chromosomes in a heterozygous individual. Such data are often called "unphased" because the allelic state of each marker is known, but the haplotype phasing is undetermined. Local phase information can be provided by, for example, cloning individual alleles before Sanger sequencing or by examining individual sequence reads generated by next-generation sequencing. Alternatively, phase information can be estimated statistically. In autotetraploids like potato, statistical phasing is much more complex and results in more phase-unknown haplotypes than for diploids (SIMKO 2004). Fortunately, once most haplotypes are known, phase information becomes less relevant, as haplotype tag SNPs can be assigned to each haplotype and explain every genotype present in a given unphased population (NEIGENFIND *et al.* 2008).

Haplotypes are essential to determine diversity parameters like allele richness and genotype composition, and to infer the biological context of alleles, because they help identify distinct variants (alleles) of genes. Other molecular marker applications, like marker-trait analysis, can be performed with either unphased SNPs or haplotype markers. However, the properties of the protein coded by an allele may depend on physico-chemical interactions involving multiple variant amino-acid sites encoded at the DNA level. In cases where these interactions influence protein function, haplotypes are of relevance in marker-trait analysis (CLARK 2004).

HIGH-THROUGHPUT SNP GENOTYPING ASSAYS

The development of high-density genotyping assays and automated genotyping tools now allows genetic screening of large populations broadly oriented towards whole genomes or targeted on specific genes (SYVANEN 2001). The large-scale SNP discovery required for these genotyping assays has generally relied on sequence variation found in libraries of expressed sequence tags (ESTs, TANG *et al.* 2006) or, more recently, on high throughput resequencing (HAMILTON *et al.* 2011). For most plant species, however, these sequence libraries contain only one or a few genotypes per species. Identified SNP markers are therefore specific to the population in which they were developed, and genotyping of broader populations will be biased towards alleles present in the original survey. This is a major problem for studies of wild or highly diverse gene pools, like that of potato. Development of SNP assays more representative of broad gene pools requires the development of efficient, cost-effective methods for genome-wide identification of large numbers of sequence variants within a large target population.

NUCLEOTIDE DIVERSITY

The rate at which nucleotide differences are observed between two randomly chosen homologous chromosomes is called the nucleotide diversity index (NEI and LI 1979). Screening more chromosomes (from more individuals) will identify more polymorphisms even while the nucleotide diversity index remains constant, thus allowing comparisons between studies using different sample sizes. Nucleotide diversity indices are reported to be 1/104 bp in maize (TENAILLON *et al.* 2001), 1/232 bp in rice (NASU *et al.* 2002) and 1/1030 bp in soybean (ZHU *et al.* 2003). Simko et al. (2006) reported in potato a nucleotide diversity index of 1/68 bp. This exceptionally high index in potato could present an obstacle for the development of SNP

genotyping assays where, to avoid hybridisation bias, constraints must be placed on sequences flanking the target SNP (SHEN *et al.* 2005).

DISCOVERY AND GENOTYPING OF DNA SEQUENCE VARIANTS BY SEQUENCING

In a genotyping-by-sequencing (GBS) approach a DNA sample is re-sequenced using first, second or third-generation sequencing technology and sequence variants are called by comparing the re-sequenced fragments to a reference sequence. As they are discovered, variants are characterized in terms of their reference (genome) sequence position and genotyped in individual samples. The discovered variants and genotyping results can be directly used for genetic analysis (VARSHNEY *et al.* 2009).

Sanger amplicon sequencing

In highly heterozygous autotetraploids such as potato, the identification and genotyping of variants is more challenging than in diploid species, because a given gene may be represented by up to four different alleles per locus in a given genotype. The genotyping method should be capable to distinguish between five zygosity classes for a bi-allelic locus: nulliplex (0:4), simplex (1:3), duplex (2:2), triplex (3:1) and quadruplex (4:0). Genotyping methods like direct Sanger amplicon sequencing have been demonstrated to be sufficiently quantitative to allow allele copy number discrimination (DE KOEYER *et al.* 2009; RICKERT *et al.* 2002; SATTARZADEH *et al.* 2006). In Sanger amplicon sequencing, sequence variants in amplicons of target genes are identified in the sequence chromatograms and then directly quantified (Figure 1).

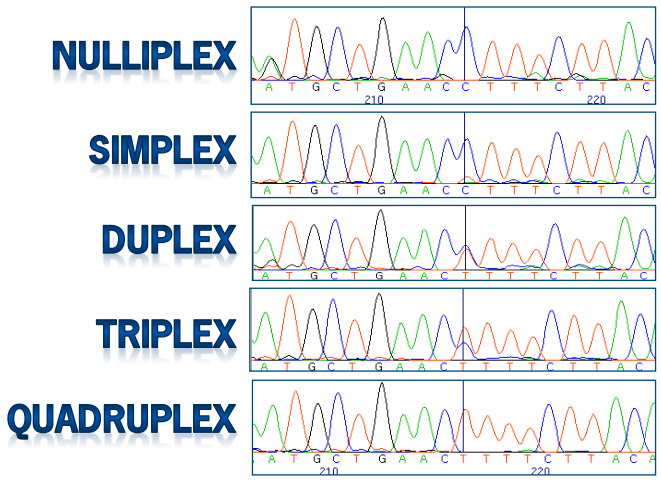

FIGURE 1. Example of genotyping-by-sequencing a gene using Sanger amplicon sequencing. Identification and genotyping of a binary C/T sequence variant in tetraploid potato; polymorphism occurs at the position indicated by the vertical line.

The Sanger amplicon strategy, although labour intensive, has been applied successful when the goal is to identification and genotyping sequence variants of a broad gene pool for a single or a limited number of target genes. For greater numbers of target genes, more effort is required to design unique primers and optimize the PCR amplification protocol for equal amplification of all alleles. Furthermore, indels can create uninterpretable sequence reads that reduce the already low throughput of Sanger-based amplicon resequencing. The speed and cost of variant discovery and genotyping of a large number of genes is thus a limiting factor for Sanger amplicon GBS.

Massively parallel sequencing

Next-generation sequencing, with its high throughput, is an effective alternative to Sanger amplicon sequencing for identifying and genotyping sequence variants in plants (ELSHIRE *et al.* 2011; NORDBORG and WEIGEL 2008). Similar to Sanger sequencing, next-generation massively parallel sequencing (MPS) can be used to directly identify and genotype sequence variants using a GBS approach (VARSHNEY *et al.* 2009). In contrast to Sanger sequencing, MPS allows resequencing of hundreds of genes, entire transcriptomes, or entire genomes with higher efficiency and more economically than ever before. MPS thus enables the sequencing, discovery, and genotyping of thousands to hundreds of thousands of markers for up to hundreds of individuals.

Many individuals can be combined in the same MPS sequencing channel using multiplex sequencing. This is achieved by the application of short (usually 4-6 bp) identifying DNA index sequences that are incorporated into the sequenced DNA fragment prior to the pooling (CRAIG *et al.* 2008; KENNY *et al.* 2011). Analysis of the first few bases of the sequence allows to assign each read to the individual it originates from.

To our knowledge, no next-generation based GBS approach has yet been applied to an autotetraploid species such as potato. This may be partly due to the fact that the sequence read depth required for accurate genotyping of polyploid species is higher than that for diploids. In humans, for example, a sequence read depth of 30-35× is considered appropriate for accurate genotyping (PELAK *et al.* 2010). Based on a binominal distribution, we estimate that accurate genotyping (p=0.95) of an autotetraploid organism, given its five possible zygosity classes, will require a sequence depth of at least 48× (Figure 2). Meanwhile, for variant discovery in an tetraploid organism, a read depth of approximately 15× has been estimated as necessary to ensure that all four alleles of the tetraploid sample are sequenced at least once (GRIFFIN *et al.* 2011). With the continuing increase in MPS data output, and the introduction of efficient complexity reduction methods to obtain the required read depths, high-throughput genotyping may be achieved most readily in autotetraploids using recently developed MPS approaches.

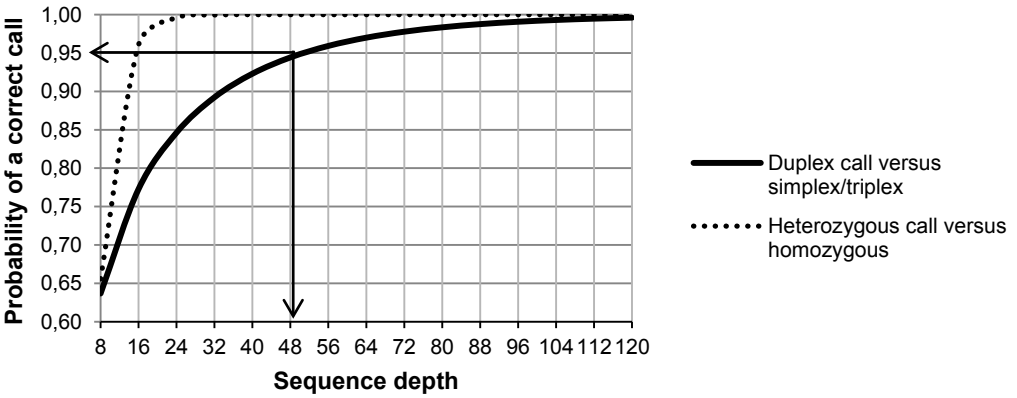

FIGURE 2. Read depth-dependent probability of correct heterozygous calls in tetraploids. Assuming balanced representation, 50% of duplex genotype reads are expected to have a given SNP at a specific position. At a read depth of eight, due to random sampling effects a duplex sample with three non-reference reads instead of four can neither be called simplex (triplex for the alternative allele) nor duplex. The probability of obtaining a proportion of non-reference reads of 3/8 or less in a duplex organism (i.e., the likelihood of correctly calling a heterozygous genotype duplex) is plotted using a binominal distribution for total read depths varying from 8 to 120×. Heterozygous genotypes can be reliably discerned from homozygous genotypes at a much lower read depths (dotted line), since only a single or a few reads with an alternative allele already affirms that the genotype is heterozygous.

Genome complexity reduction

Complexity reduction limits the portion of genetic material to be resequenced to that considered relevant for the question at hand and is often used to increase sequence depth for loci of interest. Complexity reduction strategies include cDNA libraries (GORE *et al.* 2007; HAMILTON *et al.* 2011), AFLP derived representations (VAN ORSOUW *et al.* 2007), reduced representation libraries generated by restriction- enzyme digestion and fragment selection (BAIRD *et al.* 2008; ELSHIRE *et al.* 2011), microarray-based (OKOU *et al.* 2007) or in-solution (GNIRKE *et al.* 2009; MAMANOVA *et al.* 2010) sequence capture, and additional target enrichment strategies (KISS *et al.* 2008; MAMANOVA *et al.* 2010). The use of transcriptome-based sources, such as cDNA libraries, is less suitable for GBS, since distinct alleles may have different expression levels and lead to inaccurate representation of alleles in an individual's genome. Restriction enzyme-based methods are more suitable, but these primarily target non-coding portions of the genome. In-solution sequence-capture methods like SureSelect (GNIRKE *et al.* 2009), on the other hand, make use of oligonucleotide baits designed to bind to regions of interest and can be targeted to specific sequences, whether coding or non-coding. For example, regions associated with specific traits (GNIRKE *et al.* 2009; HODGES *et al.* 2009) or all exons of a complete genome (MAMANOVA *et al.* 2010) can be targeted. Another advantage of in-solution sequence capture is its use of long 120 bp hybridization probes, which are more tolerant to polymorphisms than the shorter (typically 10-30 bp) sequences typically used in microarray-based sequence capture or as PCR amplification primers. Thus, in species with a high nucleotide diversity index like potato, these longer probes may reduce allelic bias in the complexity reduction step. Sequence-capture approaches require *a priori* availability of

sequence information from which DNA capture probes are designed. For potato, the recently sequenced genome of *Solanum tuberosum* group Phureja DM 1-3 516R44 (XU *et al.* 2011) provides an excellent resource for probe design, subsequent mapping of sequence reads, and annotation of identified variants.

MARKER-TRAIT ASSOCIATION

One of the leading motivations for the identification and genotyping of sequence variants is to achieve a better understanding of the molecular basis for naturally occurring phenotypic variation. Most of the DNA sequence variation represents neutral variation, useful for example for genetic diversity analysis, while only a small proportion is important due to its contributions to phenotypic variation in complex traits. The aim of linkage mapping is to identify genetic variants that underlie this phenotypic variation. By analysing the (co-)segregation of a phenotypic trait with certain chromosomal segments, chromosomal regions harbouring genes affecting a quantitative trait can be identified. Such chromosomal regions are referred to as quantitative trait loci (QTLs). Experimental populations for QTL analysis are generally derived from crosses of two phenotypically divergent parental genotypes. In potato, QTL mapping populations are often generated from diploid parents due to the difficulties of genotyping tetraploids. As a result, such an approach only allows monitoring of genetic variability for a maximum of four alleles (BUNTJER *et al.* 2005). Although based on the same fundamental principles of genetic recombination as linkage analysis, association mapping examines statistical association between genotypes and phenotypes in large germplasm sets of individuals with undefined relationships. In association mapping, existing plant cultivars representative of an elite genepool are usually used. Complex genetic relationships, both close and distant, among the genotypes included in an association study can therefore exist. By exploring this deeper population genealogy rather than direct offspring populations, association mapping offers three advantages over linkage analysis: much higher mapping resolution; greater allele number and broader reference population; and less research time in establishing an association. These advantages of association mapping are complementary to the strengths of linkage mapping, namely, marker efficiency and statistical power (SIMKO *et al.* 2004). The presence of subgroups with an unequal distribution of alleles within the association mapping population can generate spurious associations. That is, statistically significant associations between a marker and a phenotype may be identified in such structured populations even though the marker is not physically linked to any locus influencing the trait under consideration (FRANCIA *et al.* 2005). Fortunately, statistical approaches have been developed to account for the presence of population structure in association analysis (PRITCHARD *et al.* 2000).

Genome-wide versus candidate gene approaches

Conceptually, two different approaches for marker-trait analysis are available: a genome-wide scanning approach and a candidate gene approach. The latter applies previous knowledge about the functions of individual genes to select those most likely to influence the trait of interest. Selected candidate genes are then tested for association with the trait or phenotype. The candidate gene approach has been particularly successful for relatively simple traits

controlled by few genes, such as resistance to pests and diseases in potato. However, comprehensive screening of large potato germplasm sets has been conducted for only a few candidate genes (DE KOEYER *et al.* 2009; LI *et al.* 2005; SIMKO *et al.* 2004). In the whole-genome scan approach, also known as genome-wide association analysis (GWAS), a large number of genetic markers from across the genome is analysed to observe segregation of chromosomal segments. The strategy is to genotype enough markers across the genome so that functional alleles will likely be in linkage disequilibrium (LD) – i.e. non-randomly associated – with at least one of the genotyped markers (MYLES *et al.* 2009). D'hoop et al. (2008) were the first to explore the potential of GWAS in potato, assembling and phenotyping a germplasm panel of 430 tetraploid potato cultivars over five successive growing seasons. The panel represented a world-wide set of cultivars and progenitor lines, was complemented by breeding lines, and covered a wide range of commercial potatoes with respect to country of origin, year of release, and market segment (consumption, frying, and starch production). Genotyping was carried out using a genome-wide set of approximately 4,000 SSR and AFLP markers, followed by analysis of population genetic structure and LD for the full germplasm panel (D'HOOP *et al.* 2010; D'HOOP *et al.* 2008). Marker-trait associations were identified for traits of interest including plant maturity, tuber yield, tuber shape, fry-colour, and flesh colour (D'HOOP *et al.* 2008). The SSR and AFLP markers are unlikely, however, to be sufficiently dense to be tightly linked to sequence variants causative for trait of interests. Furthermore, since the underlying DNA sequences of the AFLP markers are unknown, translation of the anonymous markers to markers useful for marker-assisted breeding requires additional effort. This would include collecting sequence information for the markers and neighbouring genes, identifying sequence-level allelic variants of such genes, and determining which variants among these alleles affect plant phenotype.

MUTAGENESIS TO INDUCE NOVEL FUNCTIONAL SEQUENCE VARIATION

Association analysis provides only indirect (statistical) evidence of association. The genetic linkage between a specific marker and a functional target locus allele, established by association analysis, can be broken by genetic recombination. Association analysis therefore might not be sufficient to distinguish causative from phenotypically neutral polymorphisms in extended haplotype structures (ANDERSEN and LÜBBERSTEDT 2003). Sequence variants found within genes can however be used to directly study the functional relevance of those genes. Non-synonymous, splice-site, and frameshift sequence variants are particularly attractive for this goal, as they change the amino acid sequence and may therefore directly affect protein function. Such alleles can be identified in existing populations or, especially when loss-of-function alleles are desired, generated using methods like T-DNA or transposon tagging (MAY and MARTIENSSEN 2003; WALDEN 2002) and chemical mutagenesis (TILL *et al.* 2003). Ethyl methanesulphonate (EMS) mutagenesis in combination with high-throughput screening of DNA for mutations – commonly referred to as TILLING – is a rapid and cost effective method for generating novel allelic series, including loss-of-function alleles that can provide direct evidence for the functions of candidate genes related to a trait of interest.

OUTLINE OF THE THESIS

In this thesis natural and induced DNA sequence diversity in potato (*Solanum tuberosum L.*) for use in marker-trait analysis and potato breeding is assessed. The study addresses the challenges of reliable, high-throughput identification and genotyping of sequence variants in existing potato tetraploid cultivar panels using traditional Sanger sequencing and next-generation sequencing, and the application of these genetic markers. Furthermore, it explores the efficiency of ethyl methanesulphonate (EMS) mutagenesis in combination with high resolution melting (HRM) DNA screening to induce and discover novel sequence variants in potato genotypes.

In **Chapter 2** the sequence diversity in three genes of the carotenoid pathway (*CHY2*, *LCYe* and *ZEP*) is assessed in diploid and tetraploid potato genotypes using Sanger amplicon sequencing. To investigate the genetics and molecular biology of orange and yellow flesh colour in potato, association analysis between SNP haplotypes and flesh colour phenotypes is performed and the inheritance and gene expression of associated alleles is studied.

Sanger amplicon sequencing is applied in **Chapter 3** to evaluate the sequence diversity in α-Glucan Water Dikinase (*StGWD*), a candidate gene underlying a QTL involved in potato starch phosphate content. By assigning tag SNPs to haplotypes and by determining the allele copy number of identified sequence variants, we can infer the four-allele genetic composition for a large panel of tetraploid potato cultivars at this locus. This allows to estimate the average number of different haplotypes present in a single cultivar, and allows pedigree analysis to follow the allele composition over generations to confirm that the identified haplotypes are identical by descent. Association led to the identification of *StGWD* alleles causing altered starch phosphate content, which was further verified in diploid and tetraploid mapping populations containing the relevant alleles.

To scale up the discovery and genotyping of sequence variants and to make it more whole-genome oriented, **Chapter 4** reports on the next-generation sequencing of approximately 800 genes scattered over the potato genome and resequenced in 83 tetraploid potato cultivars and a monoploid reference accession. The genes targeted in this chapter are mainly single-copy genes, selected based on putative gene functions in both primary and secondary metabolic pathways, potato quality traits and biotic and abiotic stresses, and include a large set of conserved orthologous sequence genes (COSII) useful for genetic anchoring and phylogenetic studies. The accuracy of the allele copy number estimates generated by the genotyping-by-sequencing method is verified by a custom low-density SNP genotyping assay. As an example for application of genotyping-by-sequencing for genome wide association analysis (GWAS), the identified sequence variants and genotype data are tested in a marker-trait association analysis with plant maturity and tuber flesh colour. This led to the identification of alleles accounting for significant phenotypic variation in these traits.

In **Chapter 5** we apply the chemical mutagen Ethyl methanesulphonate (EMS) to diploid potato by two different treatments and screen the resulting populations for novel mutations using high-resolution melting (HRM) analysis. A pollen treatment with EMS dissolved in a

sucrose solution was found to induce mutations only at a low frequency. EMS treatment of seeds on the other hand provided a high density of novel mutations, discovered in the chimeric M_1 generation. We discovered novel sequence variants, including putative loss-of-function mutations, in six candidate genes presumed to be involved in potato starch and frying quality traits, and attempt to transfer a number of selected mutations to the M_2 and M_3 generation. The estimated mutation density of M_1 variants that are transferable to the M_2 generation (one "accessible" mutation/118-176 kb) is higher than the mutation density obtained for most other plant species, screened in the M_2 generation. The results of this chapter thus demonstrate that M_1 screening offers a good alternative to the commonly applied M_2 screening for the rapid generation of novel genetic variation at a high density, without too much complications in recovering mutations in the M_2 generation.

In the concluding **Chapter 6**, results of preceding chapters are evaluated, and the prospects of the findings for potato research and breeding are discussed.

CHAPTER 2

Identification of Alleles of Carotenoid Pathway Genes Important for Zeaxanthin Accumulation in Potato Tubers

AUTHORS
A.M.A. Wolters
J.G.A.M.L. Uitdewilligen
B. A. Kloosterman
R.C.B. Hutten
R.G.F. Visser
H.J. van Eck

Published in Plant Molecular Biology 73: 659-671, 2010

ABSTRACT

We have investigated the genetics and molecular biology of orange flesh colour in potato (*Solanum tuberosum L.*). To this end the natural diversity in three genes of the carotenoid pathway was assessed by SNP analyses. Association analysis was performed between SNP haplotypes and flesh colour phenotypes in diploid and tetraploid potato genotypes. We observed that among eleven beta-carotene hydroxylase 2 (CHY2) alleles only one dominant allele has a major effect, changing white into yellow flesh colour. In contrast, none of the lycopene epsilon cyclase (LCYe) alleles seemed to have a large effect on flesh colour. Analysis of zeaxanthin epoxidase (ZEP) alleles showed that all (diploid) genotypes with orange tuber flesh were homozygous for one specific ZEP allele. This ZEP allele showed a reduced level of expression. The complete genomic sequence of the recessive ZEP allele, including the promoter, was determined, and compared with the sequence of other ZEP alleles. The most striking difference was the presence of a non-LTR retrotransposon sequence in intron 1 of the recessive ZEP allele, which was absent in all other ZEP alleles investigated. We hypothesise that the presence of this large sequence in intron 1 caused the lower expression level, resulting in reduced ZEP activity and accumulation of zeaxanthin. Only genotypes combining presence of the dominant CHY2 allele with homozygosity for the recessive ZEP allele produced orange-fleshed tubers that accumulated large amounts of zeaxanthin.

INTRODUCTION

Flesh colour in most tetraploid potato cultivars ranges from white via cream and yellow to dark yellow. This yellow colour is caused by the presence of specific carotenoids. A small number of cultivars have red or blue/purple flesh, caused by the presence of anthocyanins.

The main carotenoids present in cultivated potato are lutein, violaxanthin, zeaxanthin and antheraxanthin (BREITBAUPT and BAMEDI 2002; BROWN *et al.* 1993; IWANZIK *et al.* 1983; NESTERENKO and SINK 2003). Beta-carotene, the precursor of vitamin A, is almost absent in S. tuberosum genotypes or closely related Solanum species (BREITBAUPT and BAMEDI 2002). Carotenoids are recognised as important health promoting ingredients of the human diet. Some have antioxidant properties, and are supposedly beneficial in preventing cancer, cardiac disease, and eye diseases (KRINSKY *et al.* 2004). Lutein and zeaxanthin are thought to be important in the human diet to prevent age-related macular degeneration (AMD) (MOELLER *et al.* 2006; SEDDON *et al.* 1994; SNODDERLY 1995). Lutein and zeaxanthin are components of the macula lutea in the human eye (HANDELMAN *et al.* 1988), protecting the retina against damaging irradiation, but they have to be replenished constantly. As humans cannot produce lutein and zeaxanthin themselves they have to be consumed by eating carotenoid-rich plant products. Lutein is present in high amounts in dark green leafy vegetables such as spinach and kale. Zeaxanthin, however, is less abundant in most vegetables (SOMMERBURG *et al.* 1998).

In tetraploid potato lutein is present in relatively large amounts, whereas zeaxanthin is present in lower amounts (BREITBAUPT and BAMEDI 2002; NESTERENKO and SINK 2003). However, some

Solanum species closely related to S. tuberosum have high zeaxanthin content (ANDRE *et al.* 2007). These are known as 'Papa Amarilla' because of their deep yellow or orange-fleshed tubers. These landraces grown by indigenous farmers in the Andean region belong to the diploid species S. stenotomum, S. goniocalyx and S. phureja (BROWN *et al.* 2007; BURGOS *et al.* 2009). Brown et al. (2007) and Brown (2008) observed a relationship between ploidy level and total carotenoid content in 38 native South American cultivars. Significantly higher mean levels of total carotenoids were observed in diploid cultivars compared with tetraploid cultivars. Morris et al. (2004) describe a diploid high carotenoid-accumulating S. phureja accession (DB375\1, or 'Inca Dawn') that predominantly contains zeaxanthin, but has a lower yield than tetraploid S. tuberosum cultivars (BRADSHAW and RAMSAY 2005). Kobayashi et al. (2008) bred a diploid potato variety with orange flesh and very high zeaxanthin content. This variety was derived from S. phureja. It would be interesting to obtain high-yielding orange-fleshed tetraploid potato cultivars to aid in the recommended daily uptake of zeaxanthin, as potatoes and potato products constitute a considerable part of the human diet in the Western world.

Yellow flesh colour in potato is mainly dependent on the presence of a dominant allele (FRUWIRTH 1912) at the Y (Yellow) locus. The Y locus has been mapped on chromosome 3 of potato by Bonierbale et al. (1988). The most likely candidate for the gene involved in yellow flesh colour is beta-carotene hydroxylase (abbreviated to Bch or CHY2) (BROWN *et al.* 2006). This gene has been mapped at the same position as the Y locus (THORUP *et al.* 2000).

Until now, the gene(s) responsible for the orange tuber flesh colour in diploid Solanum species are unknown. Brown et al. (1993) observed progeny with orange flesh colour and high levels of zeaxanthin in a hybrid population of S. phureja-S. stenotomum. They suggested that the orange phenotype was caused by a dominant Or allele at or close to the Y locus on chromosome 3 of potato. However, this was not corroborated by later research, as Brown (2008) reported. Lack of transmissibility outside the immediate 'Papa Amarilla' gene pool negated the hypothesis that the expression of Or was consistent with a strong dominant monogenic inheritance. In cauliflower an Or gene was cloned, responsible for orange-coloured curds (LOPEZ *et al.* 2008). This gene was found not to be involved in the carotenoid biosynthesis pathway, but to control chromoplast differentiation, resulting in the sequestering of large amounts of carotenoids.

This paper describes DNA polymorphisms among haplotypes of three candidate genes involved in the carotenoid pathway in monoploid, diploid and tetraploid potato genotypes, and explains the inheritance of yellow and orange potato tuber flesh colour.

MATERIALS AND METHODS

Plant materials

For sequence analyses DNA was used from five monoploid potato genotypes: 7322 (H7322 or AM79.7322, originally from G. Wenzel, Institüt für Genetik, Grünbach, Germany, see: DE VRIES *et al.* 1987; HOVENKAMP-HERMELINK *et al.* 1988), M5 and M38 (851-5 and 851-38, UIJTEWAAL *et al.* 1987), M47 and M133 (1022M-47 and 1022M-133, HOOGKAMP *et al.* 2000). DNA from 20 monoploid S. phureja and S. chacoense clones was obtained from Richard Veilleux (Blacksburg, Virginia, USA, see LIGHTBOURN and VEILLEUX 2003). DNA was isolated from eleven diploid genotypes: C (USW5337.3, HANNEMAN JR and PELOQUIN 1967), E (77.2102.37, JACOBSEN 1980), RH88-025-50 and RH90-038-21 (PARK *et al.* 2005), RH89-039-16 and SH83-92-488 (ROUPPE VAN DER VOORT *et al.* 1997; VAN OS *et al.* 2006), 87.1024/2 and 87.1029/31 (JACOBSEN *et al.* 1989), G254 (UIJTEWAAL *et al.* 1987), R5 (EJ92-6486-19, from cross 87.1024/2 x EJ91-6104-19) and 413 (transformant of interdihaploid H2260; BINDING *et al.* 1978; DE VRIES-UIJTEWAAL *et al.* 1989). Three diploid orange-fleshed genotypes were analysed: cultivars 'Papa Pura' and 'Andean Sunrise' (provided by Agrico Research BV), and S. phureja 'Yema de Huevo' (obtained from Enrique Ritter, Vitoria, Spain, see Ritter et al. 2008). Two diploid populations were analysed in which orange-fleshed progeny segregated: the CxE population (JACOBS *et al.* 1995) and the IvP92-030 population (cross G254 x SUH2293, from Ronald Hutten, Lab. of Plant Breeding, Wageningen University). Furthermore, orange-fleshed diploid genotype IvP01-84-19 from the cross 96-4622-20 x IvP92-027-9 (Ronald Hutten) was included. Additionally, a set of 225 tetraploid cultivars was used (D'HOOP *et al.* 2010; D'HOOP *et al.* 2008). In this set no genotypes with orange-fleshed tubers were present. Flesh colour of the tetraploids was determined in a field experiment in 2006 (D'HOOP *et al.* 2008). Flesh colour values were on an ordinal scale ranging from 4 (= white) to 9 (= orange) according to the Dutch Catalogue of Potato Varieties (www.nivap.nl).

DNA isolation

Genomic DNA from the monoploid and diploid genotypes was isolated from leaf tissue according to the CTAB method from Rogers and Bendich (1988). DNA from the tetraploid cultivars was isolated according to Van der Beek et al. (1992).

PCR amplification and sequencing

Amplicons for sequencing were generated from 50 ng genomic DNA template. PCR amplifications were performed in 50 µl or 25 µl reactions using 1 u of Taq polymerase, 1x reaction buffer, 200 nM dNTP and 250 nM of each primer. Standard cycling conditions were: 4 minutes initial denaturation at 94°C, followed by 35 cycles of 30 seconds denaturation at 94°C, 30 seconds annealing at 55°C and 30 seconds to 1 minute extension at 72°C. Reactions were finished by 7 minutes incubation at 72°C. For CAPS marker analysis 30 cycles were used. Most PCRs were performed with SuperTaq Polymerase buffer and enzyme (Applied Biosystems). PCR products were examined for quality on ethidium bromide-stained agarose gels. PCR products were directly sequenced on ABI377 or ABI3700 sequencers at Greenomics

(Wageningen University and Research Centre) using the dideoxy chain-termination method and ABI PRISM Reaction Kit. One or both of the amplification primers were used as sequencing primers. For SNP analysis of the CHY2 gene primers CHY2ex4F (5′-CCATAGACCAAGAGAAGGACC-3′) and Beta-R822 (5′-GAAAGTAAGGCACGTTGGCAAT-3′) were used. For SNP analysis of the LCYe gene primers AWLCYe1 (5′-AAAAGATGCAATGCCATTCGAT-3′) and AWLCYe2 (5′-GAAATACTCGGGGTACTTGAAC-3′) were used. For SNP analysis of the ZEP gene primers AWZEP9 (5′-GTGGTTCTTGAGAATGGACAAC-3′) and AWZEP10 (5′-CACCAGCTGGTTCATTGTAAAA-3′) were used. As CAPS marker for CHY2 the 308-bp CHY2ex4F + Beta-R822 PCR product was cleaved with AluI. The 163-bp fragment was indicative for the presence of CHY2 allele 3. As CAPS marker for LCYe the AWLCYe1+ AWLCYe2 PCR product was digested with SsiI, which distinguished allele 2 from alleles 1 and 3. By digesting the AWLCYe1+AWLCYe2 PCR product with HpyCH4IV allele 1 could be distinguished from alleles 2 and 3.

SNP analysis and bioinformatics

PCR reactions included a mixture of templates reflecting the different alleles in DNA samples from heterozygous diploid and tetraploid genotypes. Trace files from directly sequenced PCR products were analysed for secondary peaks, indicative for SNPs, with the Vector NTI software package from Invitrogen. Homology searches were performed at the NCBI webpage (http://www.ncbi.nlm.nih.gov/), the TGI webpage (http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/Blast/index.cgi), the SGN webpage (http://sgn.cornell.edu/), and the PGSC webpage (restricted access) (http://bacregistry.potatogenome.net/pgscreg/main.py). To study possible effects of amino acid changes on functionality of the protein the programs SIFT (Sorting Intolerant From Tolerant, see http://sift.jcvi.org/ NG and HENIKOFF 2006)) and PolyPhen (Polymorphism Phenotyping, see http://coot.embl.de/PolyPhen/) were used. Classification of the non-LTR retrotransposon sequence in ZEP allele 1 was performed at http://www.girinst.org/RTphylogeny/RTclass1.

Genetic mapping and QTL analysis in the diploid C×E population

A CxE genetic map using 94 CxE progeny was made based on an earlier version of the map (CELIS GAMBOA 2002) with additional SNP markers (ANITHAKUMARI *et al.* 2010) using mapping software Joinmap 4.0® (VAN OOIJEN 2006). QTL analysis of quantitative data was performed using the software package MapQTL® Version 5.0 (VAN OOIJEN 2004).

Cloning of ZEP promoter sequence

The promoter sequence of ZEP allele 1 was obtained by Genome Walking using the Universal GenomeWalker kit (Clontech) and the BD Advantage 2 PCR Enzyme System (BD Biosciences). DNA from diploid genotype R5 (homozygous for ZEP allele 1) was used as template. Four libraries were made using DraI, EcoRV, StuI and ScaI enzymes. The first PCR was performed with primer AP1 and gene-specific primer AWZEPGW1 (5′-TTCTGTGGAACCTTCAAATCACCGTTA-3′). The nested PCR was performed with primer AP2 and gene-specific primer AWZEPGW2 (5′-GCCCATTTTCCAAGCTCCTACAAGGTA-3′).

The DraI- and StuI-libraries yielded a 1.1-kb and 1.8-kb PCR fragment, respectively. The 1.8-kb PCR fragment of the StuI library was sequenced, and proved to contain a DraI restriction site at the expected position.

Cloning intron 1 of ZEP allele 1

To obtain the intron 1 sequence of ZEP allele 1 primers AWZEP25 (5′-CTGGCTGCATCACTGGTCAAAG-3′) and AWZEP20 (5′-TCATTCATAATTGTATCCTCCC-3′) were used. The Expand High Fidelity PCR System (Roche Applied Science) was used to obtain the 4.7-kb PCR fragment. This fragment was cloned into pGEM-Teasy (Promega). Plasmid DNA was isolated using the Promega Wizard Plus minipreps DNA Purification system. DNA from three independent colonies was first sequenced using the T7, AWZEP25 and AWZEP20 primers, and subsequently with primers designed on the obtained sequences.

Measurement of Carotenoids

From 94 progeny of the diploid CxE population carotenoids were extracted and analysed by HPLC with photodiode array (PDA) detection, according to the protocol described by Bino et al. (2005). In short, 0.5 g FW of ground and frozen tuber material was extracted with methanol/chloroform/ 1 M NaCl in 50 mM Tris (pH 7.4) in a ratio of 2.5: 2: 2.5 (v:v:v) containing 0.1% butylated hydroxytoluene (BHT). After centrifugation, the samples were re-extracted with 1 ml chloroform (+ BHT). The chloroform fractions were combined, dried under a flow of N2 gas and taken up in ethyl acetate containing 0.1% BHT. Carotenoids present in the extracts were separated by HPLC using an YMC-Pack reverse-phase C30 column and analysed by PDA detection with wavelength range set from 240 to 700 nm. Eluting compounds were identified based on their absorbance spectra and co-elution with commercially available authentic standards (neoxanthin, violaxanthin, antheraxanthin, lutein, zeaxanthin, β-cryptoxanthin, ε-carotene, α-carotene, β-carotene, ζ-carotene, δ-carotene, prolycopene and all-trans lycopene. Limit of detection was about 5 µg per 100 g FW and technical variation (6 independent extractions and analyses of the same tuber powder) was less than 8%.

In addition to the measurement of individual carotenoids yellowness of the tuber flesh of the 94 CxE progeny was determined by spectrophotometry. The same carotenoid extraction used for HPLC analysis was measured with a Perkin Elmer UV/MS Spectrometer Lambda 10. The peak area from 380 to 515 nm was determined with the UV WinLab software from Perkin Elmer using the 525 to 580 nm measurement as baseline. The peak area in the yellow spectrum is referred to as absorbance between 380 and 515 nm.

Quantitative RT-PCR

Total RNA of 23 selected genotypes of the CxE population was isolated from mature tubers as described by Bachem et al. (1998). mRNA was purified using the RNeasy mini kit (Qiagen) and reverse transcribed using the iScript cDNA synthesis kit from Bio-Rad. Relative expression level of the ZEP locus was determined by real-time quantitative reverse transcriptase PCR (qRT-PCR) on an iQ detection system (Bio-Rad) according to the Bio-rad iQ

SYBR Green Supermix protocol. The primer sequences used for the analysis were StZEP_RT_F (5`-AAGTGCCGAGTCAGGAAGCC-3`) from exon 7 and StZEP_RT_R (5`-CAAGTCCGACGCCAAGATAAGC-3`) from exon 8. Potato elongation factor 1-α (EF1α) primers were used for relative quantification (NICOT *et al.* 2005). Relative quantification of the target RNA expression level was performed using Bio-rad iQ5 analysis program.

Accession numbers

Sequence data from this article can be found in the EMBL/GenBank data libraries under accession numbers HM013963 (potato CHY2 genomic sequence), HM013964 (potato ZEP allele 1 genomic sequence) and HM013965 (potato ZEP allele 2 genomic sequence). The potato LCYe genomic sequence is available in the Third Party Annotation Section of the DDBJ/EMBL/GenBank databases under the accession number TPA: BK007065. The sequence of the PCR fragment obtained with primers AWLCYe1 and AWLCYe2 for allele 4 is available under accession number HM011105. Accession numbers for the BAC sequences are AC216345 (tomato BAC LE_HBa-11D12), AC238165 (potato BAC RH091F11), AC238104 (RH071N14), AC238398 (RH196E12), AC238240 (RH132J19) and AC215383 (tomato BAC C02HBa0104A12).

RESULTS

Allelic variation for the Beta-carotene Hydoxylase 2 gene (CHY2)

We started our investigation of the genetic requirements for orange potato tuber flesh by analysing whether orange-fleshed diploid potato genotypes contain different CHY2 alleles than white- or yellow-fleshed tetraploid cultivars. CHY2 was observed to be an important gene involved in tuber flesh colour in two separate studies performed on diploid populations (BROWN *et al.* 2006; KLOOSTERMAN *et al.* 2010). RNA expression analysis using the 44K POCI array (KLOOSTERMAN *et al.* 2008) resulted in the identification of an eQTL for yellow tuber flesh colour on potato chromosome 3 at a similar position as the CHY2 gene. Further analysis indicated that in the diploid CxE population the parent C-specific allele of the CHY2 gene was correlated with yellow flesh. This allele shows higher expression than the other two alleles segregating in the CxE population (KLOOSTERMAN *et al.* 2010). Goo et al. (2009) also observed a higher expression level of CHY2 associated with yellow flesh colour in a small number of potato cultivars.

In order to study allelic variation for the CHY2 gene in tetraploids we first determined the complete genomic sequence of the CHY2 allele in monohaploid (2n = x = 12) potato genotype 7322, as only mRNA and EST sequences were known for the potato CHY2 gene. The obtained 2,255-bp genomic sequence (Figure 1) contains seven exons. Next, we performed direct sequencing analyses of a PCR fragment obtained with primers CHY2ex4F and Beta-R822, spanning exon 4, intron 4 and exon 5 of the CHY2 gene (Figure 1). We used DNA of four additional potato monoploids, 20 S. phureja and S. chacoense monoploids, and 11 diploids. From these sequences in total eight different haplotypes could be determined (Table 1; alleles 1-7 and 11).

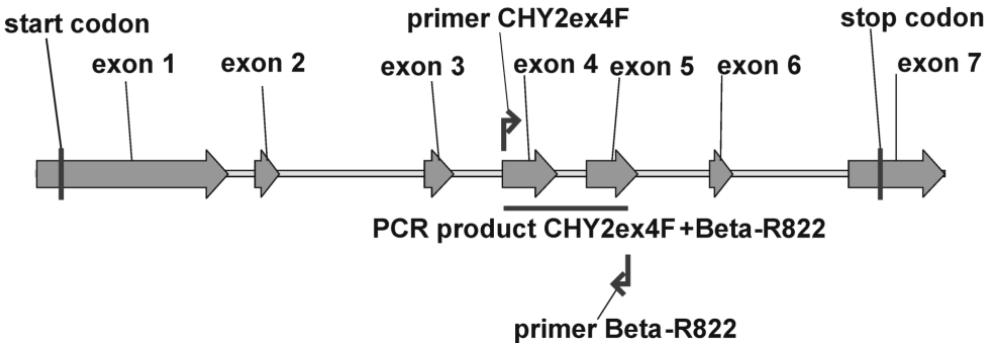

FIGURE 1. Schematic representation of the genomic sequence of the potato *beta-carotene hydroxylase* 2 (*CHY*2) gene. The 2,255-bp of allele 1 of monoploid 7322 is shown. The PCR product analysed for the presence of SNPs is indicated.

TABLE 1. Overview of haplotypes or alleles of beta-carotene hydroxylase 2 (*CHY*2). Haplotypes are based on SNP analysis using the 308-bp PCR product CHY2ex4F + Beta-R822.

	Allele										
	7322, M47, M133 [b]	M5, M38 [b]	S. phureja [c]			S.Phureja [c]					S. chacoense [c]
SNP position[a]	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
58	A	A	A	A	C	C	C	C	A	A	A
76	C	C	C	**T**[d]	C	C	C	C	C	C	C
103	C	**T**	C	**T**	C	C	C	C	C	C	**T**
105	C	C	C	C	**T**	C	C	C	C	C	C
106	A	A	A	A	A	A	A	A	A	A	**G**
123	A	A	A	A	A	A	A	A	A	**T**	A
125	A	A	A	A	**G**	A	A	**G**	A	A	A
131	A	A	A	A	A	A	A	A	A	**– (indel)**	A
136	A	A	A	A	A	A	A	A	A	**T**	A
137	C	C	C	C	C	C	C	C	C	**G**	C
142	T	T	**C**	T	T	T	T	T	T	T	T
153	**T**	A	A	A	**T**	A	A	**T**	A	A	A
163	G	A	A	A	G	G	G	G	A	G	A
171	G	G	G	G	G	**A**	**A**	G	G	G	G
244	G	A	G	A	A	G	A	G	A	G	A
250	G	A	G	A	A	G	A	G	A	A	A

[a]The SNP position is calculated from reverse primer Beta-R822
[b]7322, M47, M133, M5 and M38 are *S. tuberosum* monoploids
[c]Alleles observed in monoploid *S. phureja* and *S. chacoense* are indicated
[d]Tag SNPs or SNP combinations are in bold

We observed a correlation between presence of a single haplotype – allele 3 – and yellow flesh colour in a number of diploids. Heterozygosity for allele 3 is sufficient for yellow flesh colour, indicating this is a dominant allele. This observation corroborates the results of genetic analysis of the CHY2 (or Bch) gene by Brown et al. (2006). Furthermore, we observed that SNP 142C distinguished CHY2 allele 3 from all other haplotypes. As this SNP is unique to one haplotype it is a so-called 'haplotype tag SNP' (JOHNSON *et al.* 2001), hereafter referred to as 'tag SNP'. Allele 3 is most probably identical to the dominant allele B described by Brown et al. (2006), because sequencing of allele 3 showed the presence of allele B-specific primer sequence YellowF1. This allele is considered to be the dominant Y allele at the Yellow (Y) locus first postulated by Fruwirth (1912).

Next, we performed a SNP analysis on the DNA of a set of 225 tetraploid potato cultivars (D'HOOP *et al*. 2008). This set aims to represent the most important potato cultivars of the last 150 years in terms of acreage and/or value as progenitor, mainly from Europe, but also from the USA, Canada, and some other continents. The same PCR fragment that was analysed for the monoploids and diploids was amplified in the tetraploids. Direct sequencing was performed, which resulted in the discovery of three additional alleles (Table 2; alleles 8, 9, 10). The dosage of SNP 142C (i.e. allele 3) was determined from the sequence trace files. Dosage of allele 3 could also be estimated by using a CAPS marker assay, in which of all CHY2 alleles only allele 3 yielded a 163-bp fragment. The allele 3 dosage was related to the flesh colour value (Table 2). A flesh colour value of 5.5 or lower represents white flesh, whereas a flesh colour value higher than 5.5 indicates yellow flesh. As is shown in Table 2 presence or absence of CHY2 allele 3 is correlated with flesh colour: the group of cultivars lacking allele 3 have a mean value of 5.1 (white flesh), whereas the cultivars simplex, duplex, triplex or quadruplex for allele 3 have a mean value higher than 6 (yellow flesh). The data suggest a dosage effect of allele 3 as the quadruplex genotypes have the highest mean value. However, this group consists of only two genotypes. Although there is a clear correlation between presence of CHY2 allele 3 and yellow flesh, the intensity of the yellow flesh colour shows considerable variation within the different dosage groups, as shown in Figure 2. This was also observed by Brown et al. (2006).

TABLE 2. Relation between the number of *CHY2* alleles and tuber flesh colour.

CHY2 allele 3 dosage	*Mean tuber flesh colour value*	*Flesh colour*	*Number of genotypes*
0x (nulliplex)	5	white	69
1x (simplex)	6.3	yellow	71
2x (duplex)	6.5	yellow	46
3x (triplex)	6.6	yellow	11
4x (quadruplex)	7.4	yellow	2

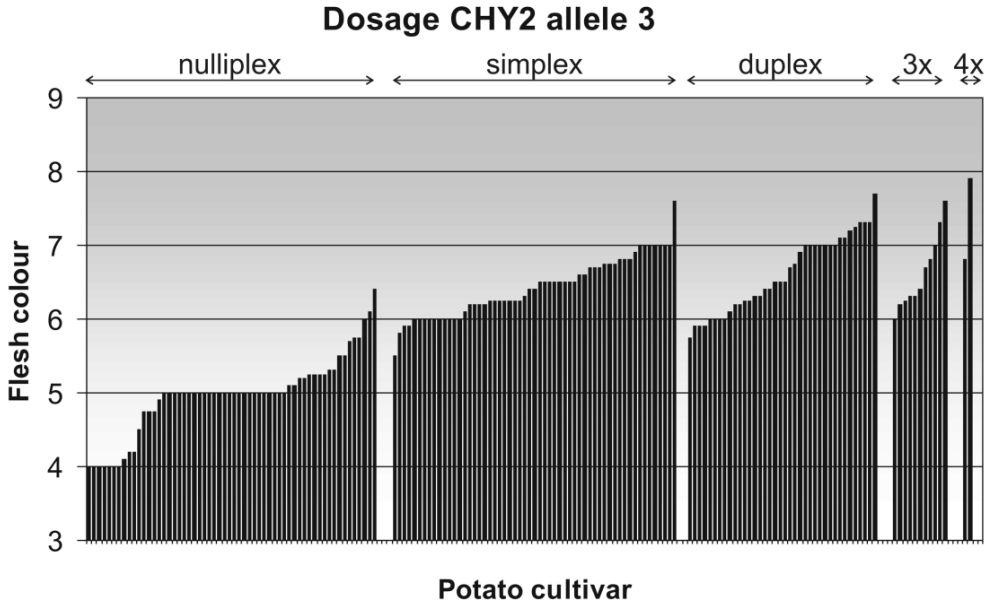

FIGURE 2. Variation in flesh colour value in classes of tetraploid potato genotypes with 0×, 1×, 2×, 3×, and 4× *CHY2* allele 3. Flesh colour value equal to or below 5.5 is considered to correspond to white flesh, values above 5.5 are indicative of yellow flesh.

To evaluate if CHY2 alleles other than allele 3 have a (small) influence on flesh colour the CHY2 allele composition was determined for 199 of the 225 tetraploid potato genotypes by analyzing tag SNPs. Large differences in allele frequency were observed. Four major alleles were observed: alleles 5 (35%), 3 (26%), 1 (20%) and 2 (13%). Four minor alleles were observed: alleles 6 (4%), 10 (2%), 8 (0.8%) and 9 (0.3%). Alleles 4, 7 and 11 were not observed in the tetraploid S. tuberosum cultivars and seem to be restricted to diploid germplasm. Minor allele 6 has been present in the Solanum tuberosum gene pool for a long time, as it is observed in nine cultivars released before 1900. Minor alleles 8, 9 and 10 seem to be novelties in the S. tuberosum gene pool, because these alleles are only present in cultivars released to the market after 1960, directly descending from backcross introgression material with late blight or cyst nematode resistance. For example, allele 10, containing an indel (1-nt deletion) in the analysed PCR fragment, is present in a number of cultivars derived from VTN 62-33-3 (1962), suggesting that this might be an S. vernei allele (see KORT *et al.* 1972).

None of the alleles 1, 2, 5, 6, 8, 9 or 10 were related to yellow flesh colour. Furthermore, none of these alleles influenced flesh colour value within the white/creamy flesh colour class, nor within the yellow flesh colour class. Thus, the variation in intensity of yellow flesh colour cannot be explained by the composition of the other CHY2 alleles in the simplex, duplex and triplex allele 3 groups.

CHY2 allele composition was determined in diploid orange-fleshed genotypes 'Papa Pura', 'Andean Sunrise', 'Yema de Huevo', IvP92-030-11 and IvP01-84-19. Although they all contained one CHY2 allele 3, they differed in the other allele, and no novel allele absent from the tetraploid gene pool was discovered. Therefore, we concluded that other genes – possibly involved in the carotenoid biosynthetis pathway –influenced the intensity of the yellow flesh colour.

Allelic variation for the Lycopene Epsilon Cyclase gene (LCYe)

The lycopene epsilon cyclase gene product is required for the synthesis of α-carotene, the precursor of lutein (see TANAKA *et al.* 2008). Silencing of the LCYe gene resulted in a significant increase in the beta-carotenoids (DIRETTO *et al.* 2006). An increase in the level of beta-carotene and zeaxanthin is expected to result in a darker yellow flesh.

To be able to analyse allelic variation for the LCYe gene in potato we needed information on the genomic sequence of the LCYe gene. Tomato BAC LE_HBa-11D12 (from chromosome 12) contains the tomato homologue of LCYe. Using this sequence four potato BAC clones from diploid genotype RH89-039-16 containing the potato LCYe homologue were retrieved from the Potato Genome Sequencing Consortium (PGSC) database: RH091F11, RH071N14, RH196E12 and RH132J19. The 7,000-bp genomic sequence of the LCYe gene contains 11 exons (Figure 3). Primers AWLCYe1 and AWLCYe2 were designed, which amplified a fragment spanning exon 7 to exon 9 of the LCYe genomic sequence (Figure 3). PCR fragments from the monoploids and diploids were directly sequenced, and 5 different alleles could be observed (Table 3). Potato BACs RH091F11 and RH071N14 contained allele 4, whereas two BACs from the homologous chromosome (RH196E12-4 and RH132J19-7) contained allele 1 of diploid genotype RH89-039-16.

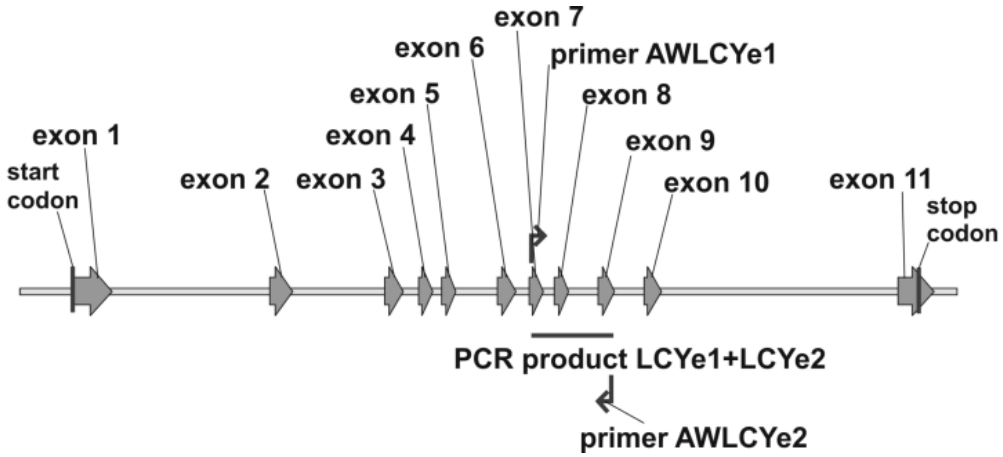

FIGURE 3. Schematic representation of the genomic sequence of the potato *lycopene epsilon cyclase* (*LCYe*) gene. A 7,000-bp sequence from BAC RH091F11 from diploid genotype RH89-039-16 is shown. The PCR product analysed for the presence of SNPs is indicated.

TABLE 3. Overview of haplotypes or alleles of lycopene epsilon cyclase (*LCYe*). Haplotypes are based on SNP analysis of the 617- or 618-bp AWLCYe1 + AWLCYe2 PCR product.

	Allele				
	7322, M47, M133, M5, M38[b]				
SNP position[a]	*1*	*2*	*3*	*4*	*5*
104	A	**G**[c]	A	A	A
115	C	C	C	C	**T**
116	**A**	G	G	G	G
121	A	**G**	A	A	A
126	**–T (indel)**	+T	+T	+T	+T
148	G	**A**	G	G	G
150	**T**	C	C	C	C
157	**A**	G	G	**T**	G
416	G	G	A	A	G
427	C	C	C	C	**T**
545	C	**T**	C	C	**T**
585	C	T	T	ND[d]	ND[d]

[a]SNP position calculated from primer AWLCYe1
[b]7322, M47, M133, M5 and M38 are *S. tuberosum* monoploids
[c]Tag SNPs or SNP combinations are in bold
[d]ND = not determined

Two alleles (allele 2 and allele 5) contain a T at SNP position 545. This nucleotide causes a change of amino acid 401 of the LCYe protein from S (serine) to F (phenylalanine). To analyse the effect of this amino acid change we consulted two software programs: SIFT and PolyPhen. According to the SIFT program an F at position 401 is not tolerated, and according to the PolyPhen program it is possibly damaging. Diploid genotype C contains LCYe alleles 2 and 3, while genotype E contains LCYe alleles 1 and 2. Thus, C and E have LCYe allele 2 in common. Therefore, 25% of the progeny of a cross between these genotypes is expected to be homozygous for allele 2. Some of the progeny of the CxE cross have a much higher flesh colour value than the yellow parent C, i.e. some have orange flesh. To investigate whether homozygosity of LCYe allele 2 leads to a difference in flesh colour 94 CxE progeny plants were analysed for their LCYe allele composition by CAPS marker assays. Data were used to localize the LCYe gene on the CxE linkage map. As expected, LCYe mapped to a position on chromosome 12 close to the STM2028 microsatellite marker and the SUS4 gene at the Southern distal end of the chromosome.

Dosage of LCYe allele 2 and dosage of CHY2 allele 3 were plotted against flesh colour value (Figure 4). This figure shows a strong correlation between presence of CHY2 allele 3 and a high value for flesh colour, whereas dosage of LCYe allele 2 does not seem to have an influence on flesh colour. This suggests that a different gene than LCYe must have an influence on intensity of yellow flesh colour.

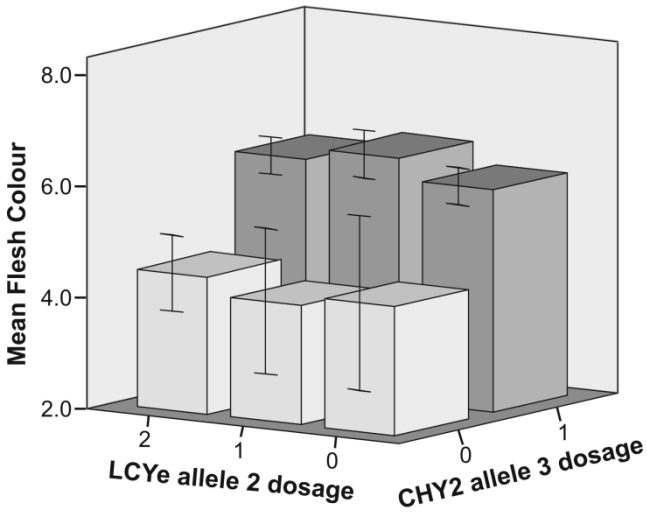

FIGURE 4. Relation between *CHY2* allele 3 dosage, *LCYe* allele 2 dosage, and phenotypic value of tuber flesh colour in the diploid CxE mapping population. Error bars ± 2 SE.

Allelic variation for the Zeaxanthin Epoxidase gene (ZEP)

Another candidate gene for orange tuber flesh in potato is the ZEP gene. ZEP is involved in the conversion of zeaxanthin into antheraxanthin, and in the conversion of antheraxanthin into violaxanthin (TANAKA *et al.* 2008). Silencing of the ZEP gene in potato resulted in transformants with higher zeaxanthin levels, and increased total carotenoid contents (RÖMER *et al.* 2002). Similarly, Morris et al. (2004) observed an inversed trend between the level of ZEP transcript level and tuber carotenoid content in a range of potato germplasm.

A genomic sequence of the tomato ZEP gene was found to be present in BAC C02HBa0104A12.1, anchored to tomato chromosome 2, which is in agreement with the map position of the pepper (Capsicum) ZEP gene on chromosome 2 (THORUP *et al.* 2000). Using the tomato BAC sequence and potato ZEP cDNA sequence DQ206629 primers were designed allowing amplification and sequencing of the complete potato ZEP gene (Figure 5). Using primer combination AWZEP9 + AWZEP10 five different alleles could be distinguished in the monoploid and diploid S. tuberosum genotypes (Table 4, alleles 1-5). The AWZEP9 + AWZEP10 PCR product, spanning exon 3 to exon 5, showed the presence of a relatively large indel in intron 4. In ZEP allele 1 a sequence of 49 bp is absent, which is present in alleles 2, 3, 4 and 5. Therefore, ZEP allele 1 could be distinguished from the other alleles by gel electrophoresis. The AWZEP9+AWZEP10 PCR product of allele 1 is 535 bp long, whereas this PCR product is 584 bp for the other alleles.

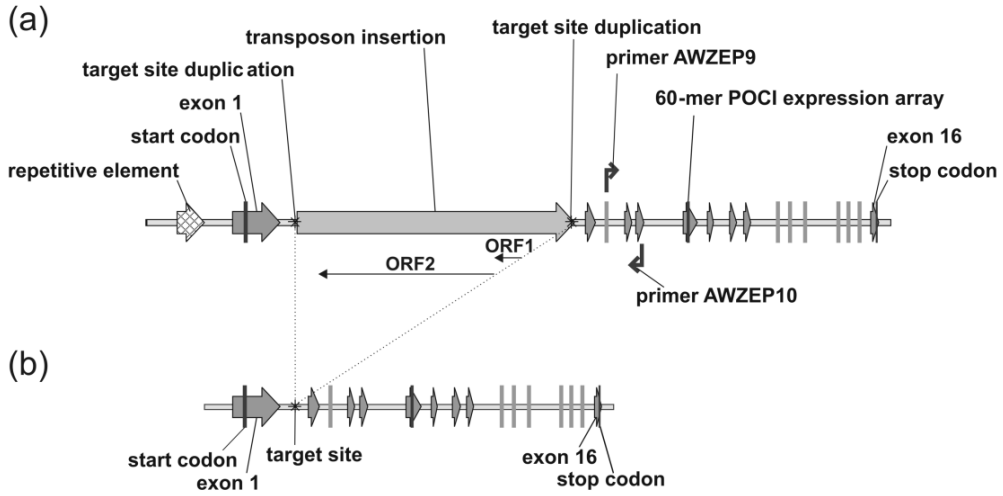

FIGURE 5. Schematic representation of the genomic sequence of the potato *zeaxanthin epoxidase* (*ZEP*) gene. (**a)** 11,000-bp sequence of allele 1, including promoter and coding region; (**b)** 6,008-bp sequence of allele 2, including promoter and coding region.

TABLE 4. Overview of haplotypes or alleles of potato zeaxanthin epoxidase (*ZEP*). Haplotypes are based on SNP analysis of the AWZEP9 + AWZEP10 PCR product.

			Allele						
	7322, M47, M5, M38[b]	*M133*[b]							
SNP position[a]	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
83	G	G	G	G	G	G	G	G	**A**[c]
99	T	T	T	T	T	**G**	T	T	**G**
113	A	A	A	A	A	**T**	A	A	**T**
154	A	A	A	A	A	**G**	A	A	**G**
165	G	G	G	G	G	G	G	**C**	G
180	A	A	A	A	A	A	A	A	**G**
209	T	T	T	T	T	T	**C**	T	T
231	T	T	T	T	T	T	T	**A**	T
375	T	T	T	T	T	T	**G**	T	T
384	A	A	**T**	A	A	A	A	A	A
389	A	A	**C**	A	A	A	A	A	A
415	A	A	A	A	**G**	A	**G**	A	A
422	**– (indel)**	C	C	C	C	C	C	T	C
426	**– (indel)**	**C**	T	T	T	T	T	T	T

[a]SNP position calculated from primer AWZEP9
[b]7322, M47, M133, M5 and M38 are *S. tuberosum* monoploids
[c]Tag SNPs or SNP combinations are highlighted

Recessive inheritance of orange tuber flesh

ZEP allele composition was determined in diploid orange-fleshed genotypes 'Papa Pura', 'Andean Sunrise', 'Yema de Huevo', IvP92-030-11 and IvP01-84-19. All five genotypes proved to be homozygous for ZEP allele 1. As these five genotypes are not closely related to each other this suggests the involvement of ZEP allele 1 in the orange flesh phenotype.

Genetic evidence for the involvement of ZEP allele 1 in orange flesh colour was obtained from cosegregation in the diploid IvP92-030 population (progeny of the cross between diploids G254 and SUH2293). This population was analysed for both CHY2 and ZEP allele compositions. Parent G254 contained CHY2 alleles 2 and 6, while parent SUH2293 contained CHY2 alleles 3 and 5. Progeny with allele combinations 2+3, 2+5, 3+6 and 5+6 were observed. Both parents G254 and SUH2293 contained ZEP alleles 1 and 2. Progeny with allele combinations 1+1, 1+2 and 2+2 were obtained in numbers compatible with the expected 1:2:1 ratio. Only progeny plants IvP92-030-9 and IvP92-030-11, containing CHY2 allele 3 and homozygous for ZEP allele 1, showed the orange-fleshed phenotype. This suggests a model in which presence of dominant CHY2 allele 3 and homozygosity for recessive ZEP allele 1 are required to obtain an orange-fleshed potato.

To investigate this further progeny of the CxE cross was analysed for ZEP allele composition. Both C and E parents contain ZEP alleles 1 and 2. A CAPS marker was developed to easily distinguish both ZEP alleles. For this, PCR product AWZEP9 + AWZEP10 was digested with the enzyme Hin6I. The PCR product of ZEP allele 1 remained undigested, whereas the PCR product of allele 2 was digested into fragments of 439 and 145 bp. Ninety-four CxE progeny were analysed with this CAPS marker. The ZEP gene was mapped in the CxE population on chromosome 2 in a similar position as the one on tomato chromosome 2. Dosage of ZEP allele 1 and dosage of CHY2 allele 3 were related to flesh colour value (Figure 6). This figure shows that homozygosity of ZEP allele 1 in combination with presence of CHY2 allele 3 results in a significantly higher mean flesh colour value. A QTL analysis for absorbance in the yellow spectrum in the CxE population resulted in a highly significant QTL on chromosome 2, on the same position as the ZEP gene (B. Kloosterman, pers.comm.).

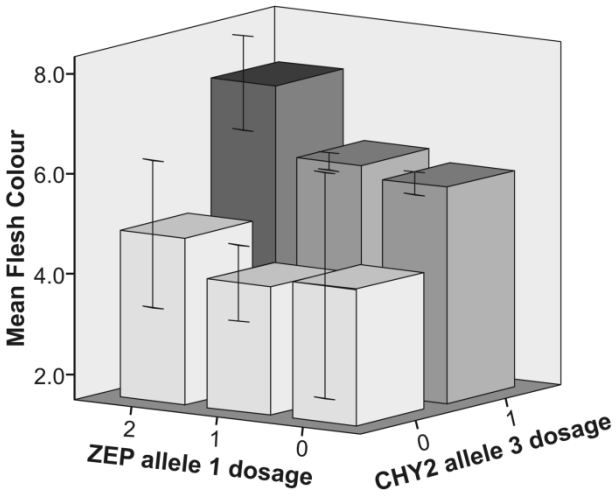

FIGURE 6. Relation between *CHY2* allele 3 dosage, *ZEP* allele 1 dosage, and flesh colour value in the diploid CxE population. Error bars ± 2 SE.

Role of ZEP Allele 1 on Zeaxanthin levels

For 88 CxE progeny the amounts of individual carotenoids were determined. A small number of progeny (10 genotypes) proved to contain relatively high levels of zeaxanthin (> 250 µg per 100 g fresh weight). These progeny invariably contained CHY2 allele 3 and were homozygous for ZEP allele 1. In Figure 7 the relation between zeaxanthin content and absorbance in the yellow spectrum is displayed for four classes of genotypes. Y or y represent the dominant or recessive CHY2 allele, respectively, and Z or z represent the dominant or recessive ZEP allele. Y is CHY2 allele 3, and z is ZEP allele 1. This figure shows that zeaxanthin only accumulates in considerable amounts in genotypes homozygous for ZEP allele 1 (zz, filled symbols). When dominant CHY2 allele 3 is present (Yyzz) the zeaxanthin level is higher than when this allele is absent (yyzz). These results suggest that orange flesh colour indicates the presence of a zeaxanthin level of more than 250 µg/ 100g fresh weight tuber.

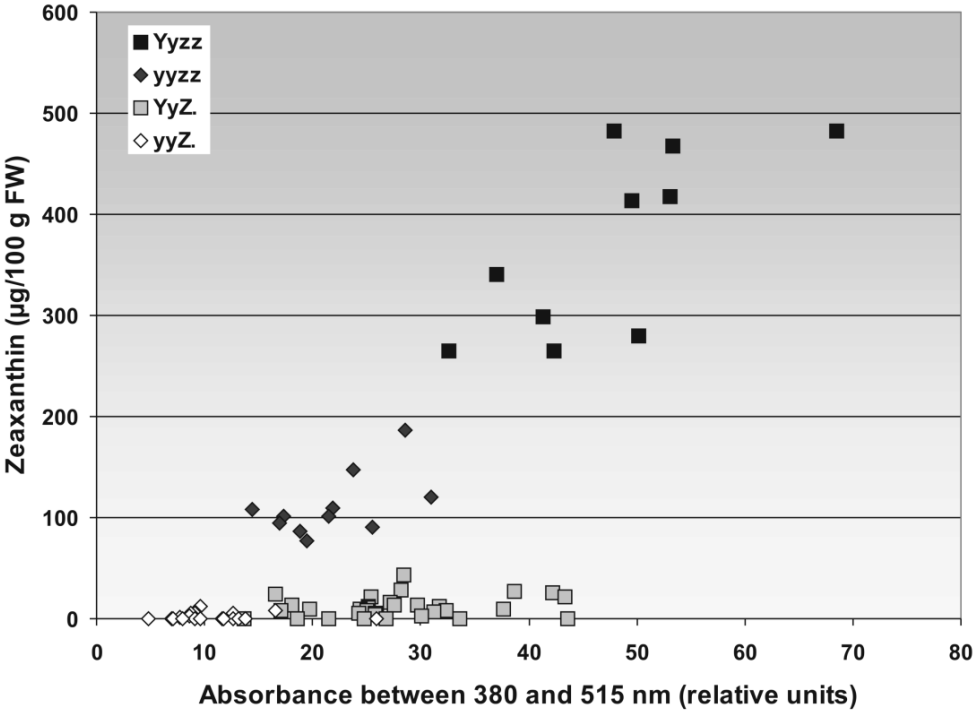

FIGURE 7. Relation between zeaxanthin content and absorbance in the yellow spectrum in the diploid CxE population. Zeaxanthin content in µg/100 g fresh weight. Four genotypic classes are indicated, with Y/y respresenting *CHY2* alleles, and Z/z representing *ZEP* alleles. Y is dominant *CHY2* allele 3; z is recessive *ZEP* allele 1.

ZEP allele 1 is expressed at a low level

Expression analysis using the 44k POCI array indicated that tuber RNA from parents C and E showed a similar level of hybridization to the ZEP-derived 60-mer oligo (B. Kloosterman, pers. comm.). Both parents are heterozygous for ZEP, containing alleles 1 and 2. Tuber RNA from CxE progeny homozygous for ZEP allele 2 showed a higher level of hybridization with the ZEP oligo than both parents, whereas tuber RNA from CxE progeny homozygous for ZEP allele 1 showed a lower level of hybridization than both parents. This may reflect a difference in homology of the ZEP alleles with the oligo. It was found that the 60-mer oligo on the POCI array (KLOOSTERMAN *et al.* 2008) is identical to a sequence in exon 6 of ZEP allele 1, while there is one mismatch with the sequence in ZEP allele 2. If the mismatch would result in a lower level of hybridization it would be expected that RNA from progeny homozygous for ZEP allele 2 would show a lower level of hybridization. However, the opposite was observed. Therefore, the array results suggest that ZEP allele 2 is expressed at a higher level than ZEP allele 1. This was confirmed by quantitative RT-PCR (Figure 8): diploid CxE progeny homozygous for ZEP allele 1 showed a significantly lower level of expression than CxE

progeny homozygous for ZEP allele 2. Heterozygous progeny displayed an intermediate level of expression. A similar observation was made by Morris et al. (2004) who found that high carotenoid-accumulating diploid S. phureja genotype DB375\1 (later renamed cultivar 'Inca Dawn') showed low expression of the ZEP gene. They observed an inverse relationship between zeaxanthin transcript level and total carotenoid content in a range of potato germplasm.

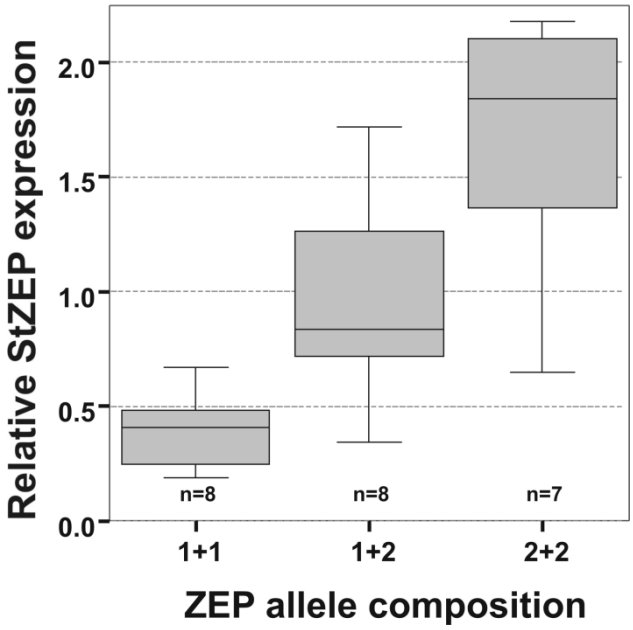

FIGURE 8. Quantitative RT-PCR of *ZEP* alleles. Relative expression level of the *ZEP* gene for CxE progeny homozygous for allele 1, heterozygous (allele 1 + allele 2), or homozygous for allele 2. *ZEP* allele 1 results in a lower expression level than *ZEP* allele.

To investigate the reason for the lower expression level of ZEP allele 1 a 1.8-kb fragment containing the promoter of this allele was obtained by Genome Walking and was sequenced. A BLASTN analysis revealed that the 5'part of this sequence contained a repetitive element (present on several chromosomes of S. tuberosum and S. lycopersicum). An analysis of cis-regulatory elements showed that the ZEP promoter contains light-regulated, phytochrome-regulated, and water stress-regulated boxes, as well as hypo-osmolarity-responsive and sugar-repression elements. Subsequently, this sequence was compared with the promoter sequence of ZEP allele 2 (as present in monoploid M133. Although a number of SNPs was observed, no obvious differences were found that could explain the different expression levels.

Next, the complete genomic sequence, including all exons and introns, was determined for both ZEP alleles 1 and 2 (Figure 5). The exon sequences were translated into protein sequences and aligned to each other. Although a number of amino acid changes were observed,

especially in the first exon, no obvious amino acid change was found predicting a non-functioning enzyme according to SIFT. Alignment of the deduced amino acid sequences of ZEP alleles 1 and 2 with ZEP protein sequences of other Solanaceous species indicated that the differences in amino acids between ZEP alleles 1 and 2 mostly occurred in the least conserved regions.

However, we observed a large difference in size in the first intron. Intron 1 in ZEP allele 2 is 389 bp in size, comparable with the 438-bp intron 1 of the tomato ZEP genomic sequence. In contrast, the size of intron 1 in ZEP allele 1 is 4,509 bp. By comparing the sequences of the first intron of ZEP alleles 1 and 2 we observed that a 4,102-bp non-LTR retrotransposon-like sequence had integrated in intron 1 of allele 1, causing a target site duplication of 18 bp. Analysis of this sequence showed the presence of two open reading frames (ORFs) in the DNA strand antisense relative to the promoter (Figure 5a). The first ORF contains an endonuclease/ exonuclease domain. The second ORF contains a non-LTR retrotransposon reverse transcriptase domain. The reverse transcriptase domain of the retrotransposon in intron 1 in ZEP allele 1 (ZEPphur) shows homology with the domains of LINE-1 like retrotransposons from other species. Analysis of the protein sequence of ORF2 at the non-LTR retrotransposon classification webpage (KAPITONOV *et al.* 2009) revealed that the retrotransposon in ZEP allele 1 belongs to the RTE clade, and is closely related to the RTE1_ZM retrotransposon from Zea mays (OBUKHANYCH and JURKA 2007). A BLASTN search of this retrotransposon sequence using the available potato genomic sequences at the PGSC webpage revealed homologous sequences on all 12 potato chromosomes, except chromosome 7. A similar search at the NCBI webpage using the high throughput genomic sequences (HTGS) database showed the presence of homologous retrotransposon sequences in several Solanaceous species, e.g. Solanum, Nicotiana, Petunia and Capsicum species. A BLASTN search using EST databases revealed that transcription of sequences homologous to the ZEPphur retrotransposon occurs.

Occurrence of ZEP allele 1 in the tetraploid potato gene pool

To determine the frequency of ZEP allele 1 in the tetraploid potato gene pool PCR with primers AWZEP9 + AWZEP10 was performed using DNA from a set of 221 tetraploid potato cultivars (D'hoop et al. 2008) and some additional cultivars. From 230 genotypes that yielded a PCR product only 5 contained ZEP allele 1, all in simplex. These were genotypes Black 1256, Prevalent (descendent from Black 1256), Producent (descendent from Prevalent), Lady Claire and Pallas (both descendents from S. phureja PHUR 71-464-7). This indicates that ZEP allele 1 is a rare allele in the tetraploid potato gene pool (frequency 0.5%).

ZEP allele composition was determined in 111 tetraploid potato genotypes not containing ZEP allele 1, by sequencing the AWZEP9+AWZEP10 PCR product. Four additional alleles were observed besides the alleles present in the monoploid and diploid genotypes (Table 4). Alleles 2 (36%), 3 (25%), 4 (14%) and 5 (20%) were major alleles, whereas alleles 6 (0.9%), 7 (0.5%), 8 (1.6%) and 9 (1.6 %) were minor alleles. The minor alleles 6 to 9 are all present in tetraploid potato cultivars released after 1960. Alleles 3 to 9 had an intron 1 of similar size as intron 1 in allele 2. Therefore none of these alleles contained the transposon insertion present in allele 1.

DISCUSSION

We conclude that homozygosity for ZEP allele 1 in the presence of dominant CHY2 allele 3 is causing the orange flesh colour phenotype, due to high levels of zeaxanthin. Furthermore, we conclude that ZEP allele 1 is a recessive allele. The accumulation of zeaxanthin does not seem to be caused by impaired function of the ZEP protein resulting from amino acid changes, as we observed no obvious amino acid changes in allele 1 compared with allele 2. Rather, the accumulation of zeaxanthin seems to result from a lower steady state mRNA level of ZEP, as determined by qRT-PCR.

Morris et al. (2004) observed an inverse relationship between the ZEP transcript level and the total tuber carotenoid content. They investigated transcript level by quantitative RT-PCR using primers designed on the basis of tomato ZEP cDNA sequence Z83835. However, DNA polymorphisms between potato and tomato can easily distort such analyses. Their forward primer contains an A at position 18 instead of G, as present in potato ZEP cDNA DQ206629 and potato EST CK278242. We observed only G at this position in our ZEP alleles, including allele 1, indicating a SNP in ZEP sequences between tomato and potato close to the 3' end of the forward primer. This mismatch may have a considerable influence on overall level of amplification in the RT-PCR experiment. We performed qRT-PCR on a diploid population segregating for ZEP alleles 1 and 2, using primers without mismatches, and observed a clear difference in expression levels between genotypes homozygous for allele 1, heterozygous (allele 1+2) and homozygous for allele 2.

A small number of SNPs was observed in the promoter sequence of allele 1 compared with the sequence of allele 2, which may explain the difference in expression level between the two alleles. However, we think it is more plausible that the difference in expression level is caused by the large retrotransposon insertion in the first intron of allele 1. This large insertion may cause inefficient splicing of the pre-mRNA into mature mRNA, or alternative splicing caused by cryptic splice sites. Hanson (1989) reported that efficient intron splicing in plants may be constrained by intron length. Ohmori et al. (2008) reported that integration of a transposon in intron 4 of rice gene DL resulted in reduced expression of the gene. Similarly, Gazzani et al. (2003) and Michaels et al. (2003) observed that weak alleles of the Arabidopsis thaliana FLC gene showing reduced expression contained a transposon in the first functional intron. A lower expression level of the zeaxanthin epoxidase gene results in the accumulation of zeaxanthin, at the expense of antheraxanthin, violaxanthin and neoxanthin (TANAKA *et al.* 2008). As zeaxanthin is relatively orange coloured (depending on concentration and milieu), while antheraxanthin, violaxanthin and neoxanthin are (light) yellow, this explains the orange flesh phenotype of the potato genotypes homozygous for ZEP allele 1 (and containing CHY2 allele 3).

Potato genotypes homozygous for ZEP allele 1 do not show the wilty phenotype as observed in Arabidopsis, Nicotiana plumbaginifolia and tomato ZEP mutants (DUCKHAM *et al.* 1991; GALPAZ *et al.* 2008; MARIN *et al.* 1996), in which synthesis of plant hormone abscisic acid (ABA), a downstream metabolite of the carotenoid pathway, is compromised. Therefore, we conclude that the reduced expression of ZEP allele 1 does not result in complete absence of ABA.

Römer et al. (2002) achieved increased zeaxanthin levels in potato tubers by genetic engineering. However, consumer acceptance of GMO cultivars is very low. Use of the natural variant ZEP allele 1 allows classical breeding for orange-fleshed potato. ZEP allele 1 probably is an S. phureja ZEP allele, as it is almost absent in the S. tuberosum gene pool, and only present in a few tetraploid potato genotypes with S. phureja in their ancestry. This means that breeding of a tetraploid potato cultivar with orange tuber flesh (with high zeaxanthin content) is a challenging task.

ACKNOWLEDGEMENTS

We wish to thank Dr. Richard Veilleux for kindly sending us DNA of 20 monoploid S. phureja/S. chacoense monoploids, Dr. Enrique Ritter for providing us with small tubers of S. phureja 'Yema de Huevo', and Agrico Research B.V. for sending us tubers of 'Papa Pura' and 'Andean Sunrise'. We thank Dr. Ric de Vos of PRI, Wageningen, for the carotenoid analysis. We thank STW for financial support of A.M.A. Wolters and J.G.A.M.L. Uitdewilligen (Grant 07926), and the EU-SOL project (PL 016214-2 EU-SOL) for financial support of B.A. Kloosterman.

CHAPTER 3

Sequence Characterization of *StGWD* Haplotypes and the Genetics of Starch Phosphate Content in Tetraploid Potato

AUTHORS
J.G.A.M.L. Uitdewilligen
A.M.A. Wolters
H.J. van Eck
R.G.F. Visser

ABSTRACT

Assessment of genetic diversity in outcrossing autotetraploid species like potato (*Solanum tuberosum*) is complex. DNA polymorphisms, often produced as unphased data, have to be combined into phased haplotypes. Once most haplotypes are known unique tag SNPs can be assigned to each haplotype, allowing full genotyping of all alleles present in an individual. Subsequently, haplotype association analysis becomes feasible. We studied the genetic diversity at a quantitative trait locus (QTL) involved in starch phosphate content by direct amplicon sequencing of the candidate gene α-Glucan Water Dikinase (*StGWD*). Sanger sequences of two *StGWD* amplicons from a global collection of 398 commercial cultivars and progenitor lines were used to identify 16 haplotypes. By assigning tag SNPs to these haplotypes, each of the four alleles present in a cultivar could be deduced. We found a nucleotide diversity value of $\pi = 16.2 \times 10^{-3}$. Between two randomly selected homologous alleles, this translated into ≈1 SNP/62 bp. A high value for intra-individual heterozygosity was observed (*Ho* = 0.765). The average number of different haplotypes per individual (*Ai*) was 3.1. Pedigree analysis confirmed that the haplotypes are identical-by-descent (IBD) and offered insight in the breeding history of elite potato germplasm. Haplotypes originating from introgression of resistance genes could be traced. Furthermore, association analysis resulted in the identification of specific *StGWD* alleles causing either an increase or decrease in starch phosphate content. These allele effects were verified in diploid and tetraploid mapping populations.

INTRODUCTION

Potato is a healthy and nutritious part of the average Western human diet, contributing carbohydrates and important amino acids and vitamins. It is, next to corn and wheat, one of the main sources of starch. Starch and its derivatives are widely employed in the manufacture of paper, textiles and adhesives, and due to their biodegradable and renewable nature they are increasingly being considered as an environmentally-friendly alternative to using synthetic additives in many other products, including plastics, detergents, pharmaceutical tablets, pesticides, cosmetics and even oil-drilling fluids (KRAAK 1992). The thermal and rheological properties of potato starch, as well as properties in processing are related to the degree of starch phosphorylation (VESELOVSKY 1940). The presence of phosphate groups in starch increases the water-binding capacity, viscosity, transparency and freeze-thaw stability of processed potato starch (CRAIG *et al.* 1989; SWINKELS 1985). Although natural starch from many plant species contains small amounts of covalently-bound phosphate, potato starch is particularly rich in phosphate (JOBLING 2004).

Starch phosphorylation is thought to play a key role in the biological breakdown of starch by affecting the physical structure of starch grains (ZEEMAN *et al.* 2007). Several genes involved in starch breakdown affect the level of starch-bound phosphate (KÖTTING *et al.* 2010) and can be proposed as candidate genes for starch phosphorylation. The most promising one is Glucan Water Dikinase (GWD), a gene first described as the R-locus in potato (LORBERTH *et al.* 1998). This single copy gene is a key enzyme in starch breakdown (EC 2.7.9.4) and catalyzes the

transfer of phosphate to the C-6 position of glucosyl residues of the amylopectin fraction (RITTE *et al.* 2006; ZEEMAN *et al.* 2007). Genetic modification of plants overexpressing this gene permits the production of high-phosphate starch (LORBERTH *et al.* 1998; RITTE *et al.* 2002). Likewise, silencing experiments have shown that potatoes with lower activity of this enzyme have a significantly lower level of starch-bound phosphate (LORBERTH *et al.* 1998). The lower amount of starch-bound phosphate decreases starch degradation and sugar accumulation in potatoes during cold storage (LORBERTH *et al.* 1998). Hence, a further reduction of phosphate in potato starch might contribute to potatoes with increased resistance to cold-sweetening. Breeding for a further increase of the phosphate content in potato starch on the other hand is highly desirable because a high natural degree of phosphorylation avoids or reduces expensive and environmentally unfriendly industrial chemical modification processes.

In the past two decades of potato research, identification of genes and markers that control the genetic variation of complex quantitative traits like starch phosphate content has mainly been done by linkage analysis in bi-parental segregating populations. These mapping populations are often developed from diploid parents that originate partly or completely from wild species. Such populations sample a maximum of four alleles in a single study and observed gene effects are often not representative of those found in elite tetraploid cultivars (SIMKO *et al.* 2004). In contrast to linkage mapping, association analysis samples a much larger number of alleles and usually cultivars with existing phenotypic information are used, representative for an elite genepool. Conceptually, there are two different approaches to identify DNA polymorphisms associated with quantitative trait loci (QTL) within an association analysis framework: a genome wide association analysis (GWAS) and a candidate gene approach. D'Hoop et al. (2008) has conducted an initial study to explore the potential of GWAS in potato by applying a genome-wide set of AFLP and SSR markers. In this study, a germplasm panel of 430 tetraploid potato cultivars was assembled and phenotyped for five successive years. The panel covered a world-wide set of cultivars and progenitor lines, complemented by breeding lines, covering the entire range of commercial potato with respect to country of origin, year of release and market segment (consumption, frying and starch industry).

In a candidate gene-based association mapping approach, genotyping is targeted to functional and positional candidate genes for the trait under consideration. Background information on the physiology and biochemistry of the trait, together with knowledge on gene function from model organisms may suggest functional involvement of the candidate gene. Additional support may be provided by positional information of QTL from linkage maps or physical maps that locate a gene to the chromosome region suspected of being involved in the trait. In potato, candidate gene-based association mapping has been conducted by re-sequencing (LI *et al.* 2005; SIMKO *et al.* 2004; WOLTERS *et al.* 2010) and more recently by High Resolution Melting (HRM) analysis (DE KOEYER *et al.* 2009). The study of De Koeyer et al. (2009) was the first to fully resolve the allelic composition of candidate genes in tetraploid genotypes. It showed that for association analysis in an autotetraploid species like potato, a genotyping technique should recognize the different alleles, as well as quantitatively asses the dosage of each allele in order to distinguish between the three different heterozygous states (simplex, duplex, and triplex). Several studies have shown that direct re-sequencing of amplicons by Sanger sequencing is

sufficiently quantitative to allow such discrimination (RICKERT *et al.* 2002; SATTARZADEH *et al.* 2006). Sanger amplicon re-sequencing of heterozygous tetraploid genotypes without cloning of the PCR product however produces unphased data. This means that even for SNPs that are only a short distance apart in the gene the primary data will not indicate how the pair of SNPs is linked in a heterozygous individual. This makes direct inference of haplotypes very complex (SIMKO 2004). However, once most haplotypes are known, unique tag SNPs can be assigned to each haplotype. In this context, unphased bi-allelic markers like SNPs can be as informative as multi-allelic molecular markers when used as "haplotype tags", that is several SNPs that tag all the detected haplotypes in a given locus. Depending on the resolution tag SNPs can be either pairwise-defined or multimarker-defined, e.g. either a single tag SNP or a combination of tag SNPs identifies a single haplotype (DE BAKKER *et al.* 2005)

Haase and Plate (1996) calculated a high heritability ($h^2 = 0.83$) for starch-bound phosphate content in potato. Werij et al. (2012) confirmed *StGWD* as a candidate gene that underlies one of the three starch phosphate QTLs in the backcross diploid C×E mapping population of potato. The QTL analysis shows three major additive QTLs on chromosomes 2, 5 and 9, each explaining approximately 20% of the observed variance. The QTL on chromosome 5 co-localizes with the *StGWD* locus. The BC_1 structure of the diploid C×E population however only allows the characterization of three GWD alleles in four possible combinations in the offspring. An unresolved questions is how the level of starch phosphorylation is influenced by *StGWD* alleles in elite potato cultivars where many more alleles, in different allelic compositions, are expected.

In this chapter we investigate whether the collective information of quantitatively scored SNPs would enable us to deduce the composition of GWD haplotypes in individual tetraploid potato cultivars. We identified 16 distinct and highly diverse haplotypes and assigned tag SNPs to each of them. With this set of tag SNPs we were able to identify the fully informative four-allele GWD configurations for almost all of the nearly 400 sampled potato cultivars. The genetic composition of the cultivars is used to identify genotypic and allelic associations with starch-bound phosphate.

MATERIALS AND METHODS

Plant material and DNA isolation

To aid haplotype identification five monoploid potato reference genotypes were used: 7322 (H7322, or AM79.7322 originally from G. Wenzel, Institut für Genetik, Grünbach, Germany), M47 and M133 (1022M-47 and 1022M-133) (HOOGKAMP *et al.* 2000) and M5 and M38 (851-5 and 851-38) (UIJTEWAAL 1987). Furthermore, DNA from nine diploid reference genotypes was used: C and E (US-W5337.3 and 77.2102.37) (HANNEMAN JR and PELOQUIN 1967; JACOBSEN 1980), 1024-2 and 1029-31 (87.1024/2 and 87.1029/31) (JACOBSEN *et al.* 1989), RH and SH (RH89-039-16 and SH82-93-488) (ROUPPE VAN DER VOORT *et al.* 1997; VAN OS *et al.* 2006), RH90 and RH88 (RH90-038-21 and RH88-025-50) (PARK *et al.* 2005), and G254 (OLSDER and HERMSEN 1976). The cultivar collection used consisted of 430 tetraploid potato cultivars and progenitor lines

(Supplementary file S1) chosen to represent a diverse range of commercial potato cultivars with respect to country of origin, year of release and market segment (D'HOOP *et al.* 2008). Genomic DNA was extracted from leaf material according to the protocol of Van der Beek et al. (1992).

StGWD gene sequence

As no genomic sequence of the potato GWD gene was available at the time we started our research we sequenced a full length genomic allele of the *StGWD* gene (Genebank JQ388473) from a BAC clone of the diploid RH89-039-16 genotype. The BAC clone (RH033J14, Genbank AC237986) was anchored to the ultra-dense SH×RH genetic map (VAN OS *et al.* 2006) and located *StGWD* to BIN37 of the upper arm of chromosome 5, between markers GP179 (BIN27) and the centromere (BIN46) and at a 12cM distance of the marker SPUD237 (BIN20) (DE JONG *et al.* 1997) (data not shown). The 16.5 kb gene contains 34 exons and encodes 1464 amino acids. In the recently sequenced *Solanum phureja* DM whole genome assembly (XU *et al.* 2011) the *StGWD* gene is located on superscaffold PGSC0003DMB000000248 of chromosome 5, and annotated as gene PGSC0003DMG400007677.

PCR amplification and sequencing

Amplification and sequencing primers (Table S1) were designed based on the consensus sequence of available genomic, mRNA and EST sequences and amplified both coding and non-coding sequence intervals of the *StGWD* gene. PCR amplicons for sequencing were generated from 50 ng genomic DNA template. Amplifications were performed in 20 µl reactions using 1 u of Taq Polymerase, 1x reaction buffer, 200 nM dNTP and 250 nM of each primer. Standard cycling conditions were: 4 minutes initial denaturation at 94°C, followed by 35 cycles of 1 minute denaturation at 94°C, 30 seconds annealing at 57°C and 40 seconds extension at 72°C. Reactions were finished by 7 minutes incubation at 72°C. PCR products were examined for quality on ethidium bromide-stained agarose gels. PCR products were directly sequenced on ABI377 or ABI3700 sequencers (Greenomics, Wageningen UR) using the dideoxy chain-termination method and ABI PRISM Reaction Kit. Forward amplification primers were used as sequencing primers. To obtain phased haplotypes PCR products of eight genotypes of the GWDex7 amplicon and six of the GWD56 amplicon were cloned in pGEM-Teasy vector (Promega) and sequenced. On average twelve cloned PCR products were sequenced for each GWD haplotype to obtain a consensus sequences.

TABLE S1. List of primers used for sequencing and genotyping *StGWD*.

Primer name	*Forward*	*Reverse*	*Optimum Tm*
GWDex7	GGAATATGAGGCTGCTCGAACT	TCTGCTCCTCCTTCTCCTTGGC	56°C
GWD56	TGAAATAAGCAAGGCTCAGGAC	ATAGTGACCTAAATCACGCAAA	55°C
GWD_G1	TCTTTGAACAGCTAGCAGAAAA	GCAGCTCTTTAACCAAAATG	57°C
GWD_G2	AACCAGGAAGTAGGAACCAG	CACACTCCCATCTCATGTTG	57°C
GWD_G34	AACAACCATCCAAACAAGGT	CCAGGACTTTTGGATAATGC	57°C
GWD_G5	TGTTCGACTGTGGACAAAAC	AGGGTTGCTATGTGAATGGT	57°C
GWD_G6	TAATGGTGATCCATTTGCAG	CAGAGAGTGCGAGATTTTCA	57°C
GWD_G7	TCAAGCTCTTTCAATGTCCA	AATCCTTCCTTCCTTTGCTT	57°C
GWD_end_2	GGATGAGGAGGAAAAAGTTG	TGCAATACATAATGCGTGTG	57°C
GWD_HRM_#1	CTTGAGCTTGAGAAAGGCAT	CAAAGTCTCTCTTCTTTCTTTGGAT	60°C

Sequence variant detection and analysis

Alignment and quality scoring was done using the Staden software package (STADEN 1996). Sequence variations (SNPs and short Indels) were detected using NovoSNP (WECKX *et al.* 2005). The allele copy number of SNPs was scored using both the Data Acquisition & Data Analysis software DAx7.1 (Van Mierlo Software Consultancy) and manual scoring. For nucleotide diversity and phylogenetic analysis the consensus haplotype sequences were compared with one another and with *S. lycopersicum*-derived sequences using MEGA 4 (TAMURA *et al.* 2007) and TREECON (VAN DE PEER and DE WACHTER 1994) software. Similarity between each pair of sequences was calculated on the basis of percentage identity and tree construction was performed using the Neighbor-joining method. To estimate gene frequencies a program for the analysis of autotetraploid genotypic data, AUTOTET (THRALL and YOUNG 2000) was used. The following statistics were calculated to describe the levels of genetic diversity: *Ai*, the average number of alleles per individual at a locus; *Ho*, the observed heterozygosity; and *He*, the expected heterozygosity. In order to compare the genotype proportions with those expected under Hardy–Weinberg equilibrium the mean fixation index (*F*) was calculated and the chi-squared (χ^2) test was used to evaluate deviations of *F* from zero. Pedigree information was collected from the potato pedigree database (VAN BERLOO *et al.* 2007) and inspected for abnormalities in *StGWD* allele transmission using Pajek (DE NOOY *et al.* 2005) and Cytoscape (SHANNON *et al.* 2003).

High resolution melting analysis

Amplicons for HRM genotyping were generated from 15 ng genomic DNA template. PCR amplifications were performed in 10 µl reactions using 2 µl of F-524 Phire™ 5× reaction buffer (Finnzymes), 0.1 µl Phire™ Hot Start DNA Polymerase (Finnzymes), 1 µl LCGreen™ Plus+ (BioChem) and 0.25 µl of 5 mM primers. PCR and heteroduplex formation were performed using the following conditions: 94°C, 2 minutes; 40 cycles, 94°C, 5 seconds; fragment-dependent Tm, 10 seconds; 72°C, 10 seconds; a denaturation step of 30 seconds at 94°C and renaturation by cooling to 30°C. Amplicons were genotyped using the LightScanner® System (Idaho Technology).

Phenotypic data collection

The tetraploid genotypes of this study were grown in two years and starch was isolated from both years. The phosphate content of starch was however analyzed for only one year and for a subset of 207 genotypes, because the assay is laborious. Starch phosphate measurements of individual samples were repeated in triplicate. For this measurement, approximately 20 mg starch (dry weight) was added to 250 µl 70% $HClO_4$ and heated at 250°C for 25 minutes. 50 µl 30% H_2O_2 was added and the mixture was heated at 250°C for another 5 minutes. After cooling down the volume was increased to 2 ml by adding H_2O. 100 µl of the sample was pipetted into a 96-well microtiter plate and 200 µl of color reagent (0.75% $(NH_4)6Mo_7O_{24}.4H_2O$, 3% $FeSO_4.7H_2O$ and 0.75% SDS dissolved in 0.375 M H_2SO_4) was added. After incubation for 10 minutes at room temperature, the absorbance was measured at 750 nm, and compared to the absorption of a calibration curve to determine the sample PO_4 concentration in nmol PO_4/mg starch.

Association analysis

For the analysis of phenotypic data and marker-trait association SPSS (IBM) was used. The association analysis was performed using a linear mixed model. The multivariate model was arranged to simultaneously assess the significance of all haplotype effects. Copy number of each of the haplotypes were modeled as fixed effects and haplotypes A_1 to A_5 were modeled as nested factors of grouped haplotype A. Variance components were estimated by the REML method. A general linear model was applied to estimate the explained phenotypic variance of associated haplotypes. For this, copy number of the haplotype within each clone was tested separately.

RESULTS

Sequence diversity and haplotype analysis

A panel of five monoploid and nine diploid potato accessions was selected to gain an initial insight into *StGWD* nucleotide polymorphism among *S. tuberosum* clones. Seven PCR amplicons were Sanger sequenced and assessed for single locus amplification, SNPs and Indels. Amplicons derived from the monoploid accessions had sequence chromatogram peaks representing a single haplotype. Amplicons of diploid accessions displayed double chromatogram peaks at discrete nucleotide positions as expected for heterozygous accessions. Of the seven different amplicons, three amplicons showed no indel polymorphisms. Indel polymorphisms can result in undecipherable sequence chromatograms. Two amplicons were selected to identify SNPs and haplotypes in a broader panel of 430 tetraploid potato cultivar and progenitor lines. The GWDex7 amplicon (627 bp) includes a large part of the gene region from exon 8 to exon 9. The GWD56 amplicon (606 bp) covers exon 15 to exon 17 (Figure 1).

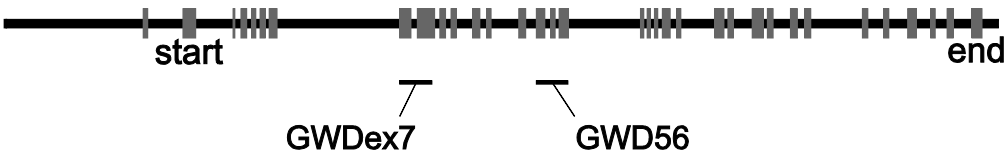

FIGURE 1. Gene model of StGWD genomic sequence. The 19,188 bp genomic sequence from genotype RH89-039-16 contains 34 exons (grey boxes) and 4,392 bp of coding sequence. The GWDex7 and the GWD56 amplicons used for re-sequencing and genotyping are indicated.

For the tetraploid accessions, high quality sequence chromatograms with an average read length of 523 bp were generated for 398 cultivars. A small number of accessions showed low quality chromatograms in repetitive runs. In the approximately 1 kb of accessible DNA sequence of the two amplicons, 81 polymorphisms were detected and quantitatively scored. Four of these polymorphisms were of multi-allelic (three tri-allelic, one tetra-allelic) nature and 77 of bi-allelic nature. The average number of polymorphic sites – which ignores the fact that polymorphisms co-segregate in haplotype blocks – was 1 polymorphism /12 bp.

Using the sequence information of both monoploid and diploid accessions seven initial haplotypes were inferred (A_1, A_2, A_3, B, C, D, F). Three haplotypes (A_1, A_2, B) were observed

among the five monoploid accessions. Haplotypes of the diploids were deduced by subtracting already identified haplotypes from the sequence chromatograms. Haplotypes of the tetraploid potato germplasm collection could not be directly inferred from the unphased sequence chromatograms due to the highly heterozygous state and high SNP frequency. Putative haplotype models for these accessions were deduced by identifying sets of co-segregating SNPs. For this we calculated the squared correlation coefficient (r^2) between the copy numbers of all polymorphisms. Co-segregating polymorphisms were assigned to putative haplotypes and novel haplotypes were identified by sequencing cloned amplicons of a number of corresponding potato accessions. All polymorphisms were assigned to 16 verified haplotypes, and haplotype-specific tag SNPs could be identified (Table 1).

TABLE 1. Phased GWD haplotypes for amplicon (A) GWDex7 and (B) GWD56. Positions are relative to the start of the amplicon. Haplotype-defining tag SNPs are color-coded; Dark grey bases indicate SNPs which tag a single haplotype, light grey bases indicate SNPs shared by multiple haplotypes. Deletions in haplotypes are shown as asterisks. The last three lines in the tables indicate the amino acids and their codon position in the reference sequence and non-synonymous changes. The non-coding SNPs in introns are indicated by missing codon positions (-). Tag SNPs used for copy number estimation are shown in bold.

(a) Amplicon GWDex7

GWDex7	*82*	*115*	*137*	*139*	*151*	*155*	*162*	*163*	*209*	*228*	*237*	*251*	*261*	*265*	*268*	*271*	*283*	*289*	*292*	*324*	*398*	*403*	*407*	*418*	*419*	*425*	*438*	*448*	*459*	*460*	*470*	*477*	*518*	*534*	*551*	*568*	*579*	*580*	*581*	*598*	*600*
A_1	A	G	G	A	C	G	C	A	C	A	T	A	**C**	T	G	T	**T**	T	T	G	C	A	G	A	G	A	A	C	G	T	C	G	G	C	A	T	C	C	G	A	C
A_2	A	G	G	A	C	G	C	A	C	A	T	A	T	T	G	T	**T**	T	T	G	C	A	G	A	G	A	A	C	G	T	C	G	G	C	A	T	C	C	G	A	C
A_3	A	G	G	A	C	G	C	A	C	A	T	A	T	T	G	T	**T**	T	T	G	C	A	G	A	G	A	A	C	G	T	C	G	G	C	A	T	C	C	G	A	C
A_4	A	G	G	A	C	G	C	A	C	A	T	A	T	T	G	T	**T**	T	T	G	C	A	G	A	G	A	A	C	G	T	C	G	G	C	A	T	C	C	G	A	C
A_5	A	G	G	A	C	G	C	A	C	A	T	A	T	T	G	T	**T**	T	T	G	C	A	G	A	G	A	A	C	G	T	C	G	G	C	A	T	C	C	G	A	C
B	A	G	G	T	C	G	C	A	C	A	C	A	T	T	C	C	G	T	G	G	C	A	G	A	G	A	A	T	C	**A**	T	G	T	C	A	T	C	C	A	A	C
C	C	G	A	T	C	G	C	A	C	A	C	T	T	T	C	C	G	T	G	G	C	A	G	A	G	A	**G**	T	C	T	T	G	G	C	A	T	G	C	A	A	C
D	A	G	G	T	C	G	C	A	C	A	C	A	T	T	C	C	G	T	G	G	C	A	G	**C**	G	A	A	T	C	T	T	G	T	A	A	T	C	C	A	A	C
E	A	A	G	T	C	G	T	A	C	A	C	A	T	T	C	C	G	T	G	**A**	C	A	G	A	G	A	A	T	C	C	T	G	T	A	G	T	C	C	A	A	C
F	A	G	G	T	C	G	C	A	C	**G**	C	A	T	T	C	C	G	T	G	G	C	A	G	A	G	C	A	T	C	T	T	G	T	C	A	C	C	T	T	A	A
G	A	G	G	T	C	G	C	A	C	A	C	A	T	T	C	C	**A**	C	G	G	C	A	G	A	G	A	A	T	C	T	T	G	T	C	A	T	C	C	A	A	C
H	A	G	G	T	C	T	C	A	C	A	C	A	T	T	C	C	G	T	G	G	**A**	A	G	A	G	A	A	T	C	T	T	T	T	C	A	T	C	C	A	G	A
I	A	G	A	T	C	G	T	A	C	A	C	A	T	**C**	C	C	G	T	G	G	C	A	G	A	G	A	A	T	C	T	T	G	G	C	A	T	C	C	A	A	C
J	A	G	G	T	C	G	C	A	C	A	C	C	T	T	C	C	**C**	T	G	G	C	A	G	A	G	A	A	T	C	T	T	G	T	C	A	T	C	C	A	A	C
K	A	G	G	T	C	G	C	A	C	A	C	A	T	T	C	C	G	T	G	G	C	A	G	**C**	**C**	A	A	T	C	T	T	G	T	A	A	T	C	C	A	A	C
L	A	G	G	T	**T**	G	C	G	T	A	C	A	T	T	C	C	G	T	G	G	C	G	A	A	G	A	A	T	C	T	T	G	T	C	A	T	C	C	A	A	C
Amino-acid	T	E	N	N	D	A	A	A	P	-	-	-	-	-	-	-	-	-	-	-	T	D	E	K	K	T	E	V	L	L	S	A	Q	P	V	L	A	A	A	K	L
Codon position	273	284	292	292	296	298	300	300	316	-	-	-	-	-	-	-	-	-	-	-	339	341	342	346	346	348	353	356	360	360	363	366	379	385	390	396	400	400	400	406	407
AA change	-	-	D	E	-	S	V	V	S	-	-	-	-	-	-	-	-	-	-	-	-	G	-	T	T	-	K	A	V	Q/P	-	S	H	T	-	S	P	L	L	R	M

(b) Amplicon GWD56

GWD56	85	134	185	187	195	199	210	211	212	215	233	246	253	256	257	265	277	289	298	352	384	391	402	405	407	408	410	411	417	418	423	425	441	489	507	545	555	577	578	579
A_1	G	C	T	**A**	G	C	C	A	G	G	T	G	G	C	T	**T**	A	G	G	G	A	A	C	G	A	G	A	G	C	G	A	T	G	C	A	G	G	A	A	T
A_2	G	C	T	**A**	G	C	C	A	G	G	T	G	G	C	T	**T**	A	G	G	G	A	A	C	G	A	G	A	G	C	G	A	T	G	C	A	G	G	A	A	T
A_3	G	C	T	**A**	G	C	C	A	G	G	T	G	G	C	T	A	A	G	G	G	A	A	C	G	A	G	A	G	C	G	A	T	G	C	A	A	G	A	A	T
A_4	G	C	T	**A**	G	C	**T**	A	G	G	T	G	G	C	C	A	A	G	G	G	A	A	C	G	A	G	A	G	C	G	A	T	G	C	A	A	G	A	A	T
A_5	G	C	T	**A**	G	C	C	A	G	G	T	G	G	C	T	A	A	G	G	G	A	A	C	G	A	G	A	G	C	G	A	T	G	C	A	A	G	A	A	**C**
B	G	C	T	T	C	C	C	A	G	**A**	T	G	G	C	T	A	C	G	G	G	A	A	C	G	A	G	A	**A**	C	A	A	T	A	C	T	A	A	G	G	T
C	G	A	T	T	C	C	C	A	G	G	T	G	G	C	T	A	A	A	G	G	A	A	**A**	G	A	A	A	G	C	A	A	G	G	A	A	A	A	G	G	T
D	G	C	**C**	T	C	C	C	A	G	G	T	G	G	C	T	A	C	G	G	G	A	A	C	G	A	G	A	G	C	A	A	T	A	C	A	A	A	G	G	T
E	G	C	**C**	T	C	**G**	C	*	*	G	T	G	G	C	T	A	C	G	G	G	A	A	C	G	**T**	G	**T**	G	C	G	A	T	A	C	A	A	A	G	G	T
F	G	C	T	T	C	C	C	A	G	G	**C**	G	G	C	T	A	**T**	G	G	G	**G**	A	C	G	A	G	A	**A**	C	A	A	T	G	C	A	A	A	A	A	T
G	G	C	T	T	C	C	C	A	G	**A**	T	G	G	C	T	A	C	G	G	G	A	A	C	G	A	G	A	G	C	A	A	T	A	C	**T**	A	A	G	G	T
H	C	C	T	T	C	C	C	A	G	G	T	G	**A**	A	T	A	C	G	G	G	A	**T**	C	G	A	G	A	G	**T**	A	A	T	A	C	A	A	A	G	G	T
I	G	A	T	T	C	C	C	A	G	G	T	**A**	G	C	T	A	A	A	G	G	A	A	C	G	A	G	A	G	C	A	A	T	A	C	A	A	A	G	G	T
J	G	C	**C**	T	C	C	C	A	G	G	T	G	G	C	T	A	C	G	G	G	A	A	C	**T**	A	G	A	G	C	A	A	T	A	C	A	A	A	G	G	T
K	G	C	**C**	T	C	C	C	A	G	G	T	G	G	C	T	A	C	G	G	**A**	A	A	C	G	A	G	A	G	C	A	**G**	T	A	C	A	A	A	G	G	T
L	G	C	T	T	C	C	C	A	G	G	T	G	G	C	T	A	C	G	**C**	G	A	A	C	G	A	G	A	G	C	A	A	T	A	C	A	A	A	G	G	T
Amino-acid	R	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	K	K	M	V	-	-	-	-	-	-	-	-	-	-	-	-	L	T	I	H	P	G	G	G
Codon position	653	679	-	-	-	-	-	-	-	-	-	-	-	-	-	-	681	685	688	706	-	-	-	-	-	-	-	-	-	-	-	-	711	727	733	746	749	757	757	757
AA change	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	N	-	I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	R	-	N	N	N

The quantitatively scored tag SNPs were used to assign an allele copy number and genotype composition to each cultivar. In case haplotypes contained multiple tag SNPs tagging the same haplotype the best quantifiable SNP was selected for copy number estimate. When a haplotype contained no unique tag SNP the allele copy number was inferred by subtracting the copy number of already tagged allele(s) from the "tag" SNP shared by the alleles (Table 2). For the GWDex7 amplicon identical-in-state haplotypes A_2, A_3, A_4 and A_5 and haplotype D were multi-marker defined. For the GWD56 amplicon haplotypes A_1 and A_2 were identical-in-state and there were four multi-marker defined haplotypes. The allele copy numbers found for the haplotypes of the GWDex7 amplicon invariably matched the allele copy numbers in the GWD56 amplicon. Using the selected tag SNPs it was possible to assign a four-allele genotype to 384 (96%) of the tetraploid potato cultivars (Supplementary file S1).

TABLE 2. GWD haplotype tag SNPs. All SNPs were quantitatively scored and used for copy number estimation, but only indicated SNPs were used to estimate haplotype copy number and to detect the full four-allele genotype of the tetraploid cultivars. For a cultivar re-sequenced successfully in only one amplicon, either GWDex7 or GWD56, the A haplotypes can be identical-in-state to each other. Some haplotypes are without unique haplotype tag SNP and are multi-marker defined. E.g. haplotype D in the GWDex7 amplicon is defined by SNP418C - Allele K (= SNP419C).

	Amplicon tag SNP	
GWD Haplotype	*GWDex7*	*GWD56*
A_1	SNP261C	SNP265T
A_2	SNP283T - Allele A_1	SNP265T
A_3	SNP283T - Allele A_1	SNP187A - Allel A_4 - Allel A_5
A_4	SNP283T - Allele A_1	SNP210T
A_5	SNP283T - Allele A_1	SNP579C
B	SNP460A	SNP215A - Allel F
C	SNP438G	SNP402A
D	SNP418C - Allele K	SNP185C - Allel E - Allel J - Allel K
E	SNP324A	SNP199G
F	SNP228G	SNP384G
G	SNP283A	SNP215A - Allel B
H	SNP398A	SNP253A
I	SNP265C	SNP246A
J	SNP283C	SNP405T
K	SNP419C	SNP352A
L	SNP151T	SNP298C

To investigate the sequence similarity between the detected haplotypes, a Neighbor-joining dendrogram was constructed using amplicon sequences of *S. lycopersicum* as out-group (Figure 2). Over the 1 kb of DNA sequence of the two amplicons, the tomato haplotype was to a high degree similar (95.4-96.5%) to the haplotypes observed in the potato germplasm set. Sequence similarity between the 16 potato haplotypes ranged from 96.8% to 99.9%. Distance

between the two most distant potato haplotypes (A_1 and E) approached the sequence divergence observed between potato and tomato.

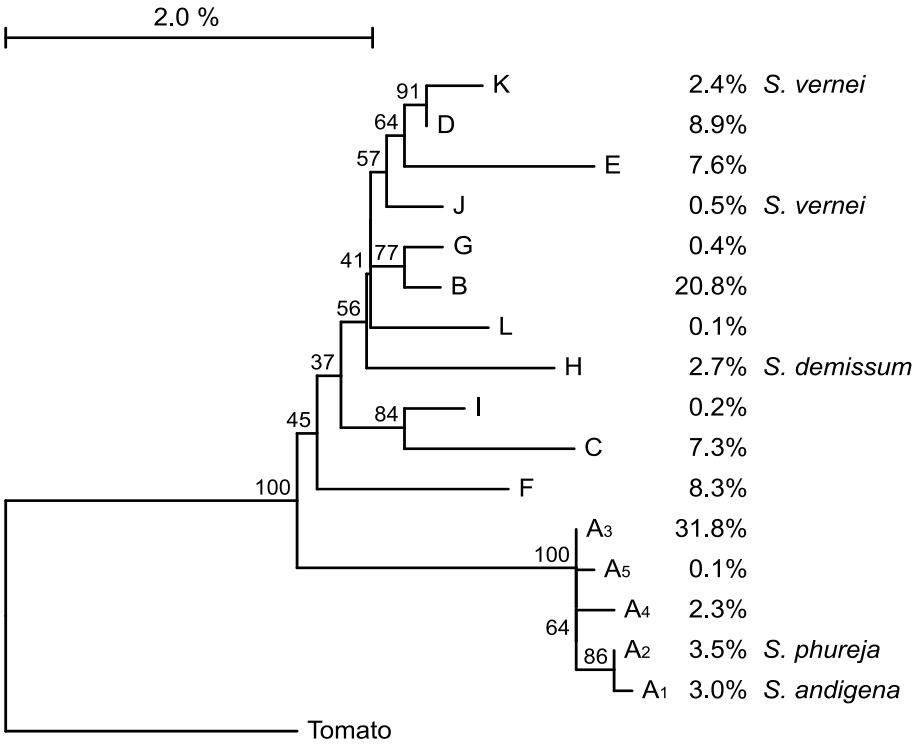

FIGURE 2. Dendogram of the 16 GWD haplotypes. The distances were computed using the Jukes-Cantor method and the tree inferred using the Neighbor-joining method. The percentage of replicate trees in which the associated haplotypes clustered together in the bootstrap test (1000 replicates) are shown next to the branches. For each allele the frequency and – when identified – the source is given. The tomato sequence was used as out-group to root the tree.

The estimated nucleotide diversity between the 16 potato haplotypes was $\pi = 18.5 \times 10^{-3}$ and translated into an average SNP diversity of ≈1 SNP/54 bp ($1/\pi$). At the protein level, the analyzed haplotypes included 302 codons of five exons. Of those codons 36 showed polymorphisms, causing 24 non-synonymous changes and 15 synonymous changes. No well-defined dysfunctional mutations such as stop codon, splicing site or frame shift mutations were found. For estimates of nucleotide diversity at the population level, the frequencies at which haplotypes occurred was considered. Six haplotypes had an allele frequency above 5%, and 10 had a frequency below 5% (Table 3). In the sampled population of 398 cultivars, we found a population frequency adjusted nucleotide diversity value of $\pi = 16.2 \times 10^{-3}$. Between two randomly selected homologues alleles, this translated into ≈1 SNP/62 bp.

Pedigree analysis

To verify the that the identified haplotypes were identical-by-descent and to identify putative sources of the haplotypes we performed a pedigree analysis. For 218 fully genotyped cultivars at least one parental cultivar had also been genotyped, and for 56 of these both parents were genotyped. The pedigree-based relationships and the four-allele GWD genotypes of the cultivars are shown in (Figure S1). For 22 out of the 218 genotyped parents/offspring pairs a mismatch was observed. In 12 occasions the mismatch repeatedly involved the parental genotypes of the cultivars AM 78-3704, Sirtema, Early Rose and Patersons Victoria. For several of the alleles, the putative source of the allele was found (Table 3). Haplotypes A_1, H, I, J and K were found to be relatively new in the analyzed genepool. Haplotype G was found only in five heirloom potato cultivars. Other haplotypes were present in both ancient and new potato cultivars (Supplementary file S1).

TABLE 3. Allele-frequencies of GWD haplotypes in the collection of ~400 sequenced tetraploid potato cultivars and breeding lines. Five haplotypes have an allele frequency below 1% and only six haplotypes have an allele frequency above 5% (major alleles). The haplotype A group contains the minor alleles A_1, A_2, A_4 and A_5 and the common allele A_3. By examining potato pedigree data the putative donor of some of the minor alleles is identified.

GWD Haplotype	*Allele count*	*Allele Frequency*	*Possible sources*
A_1	46	3.00%	Observed in descendants of *S. andigena* clone CPC 1673 used as donor of resistance against *Globodera rostochiensis*
A_2	49	3.50%	Observed in *S. phureja* genotype DM1-3 516R44
A_3	464	31.80%	.
A_4	33	2.30%	.
A_5	2	0.10%	Observed only in Lenape and Golden Wonder
B	331	20.80%	.
C	117	7.30%	.
D	139	8.90%	.
E	121	7.60%	.
F	129	8.30%	.
G	6	0.40%	Observed in heirloom cultivars
H	43	2.70%	Observed in progeny of *S. demissum* introgression clone USDA 96-56 used as donor for R1 resistance against *Phytophthora infestans*
I	3	0.20%	Observed in Astarte and its descendants
J	8	0.50%	Observed in progeny of *S. vernei* introgression clone VE 66-295
K	39	2.40%	Observed in progeny of *S. vernei* introgression clone VTN 62-33-3, donor of resistance against *Globodera pallida* Pa2
L	1	0.10%	Observed only in Hindenburg

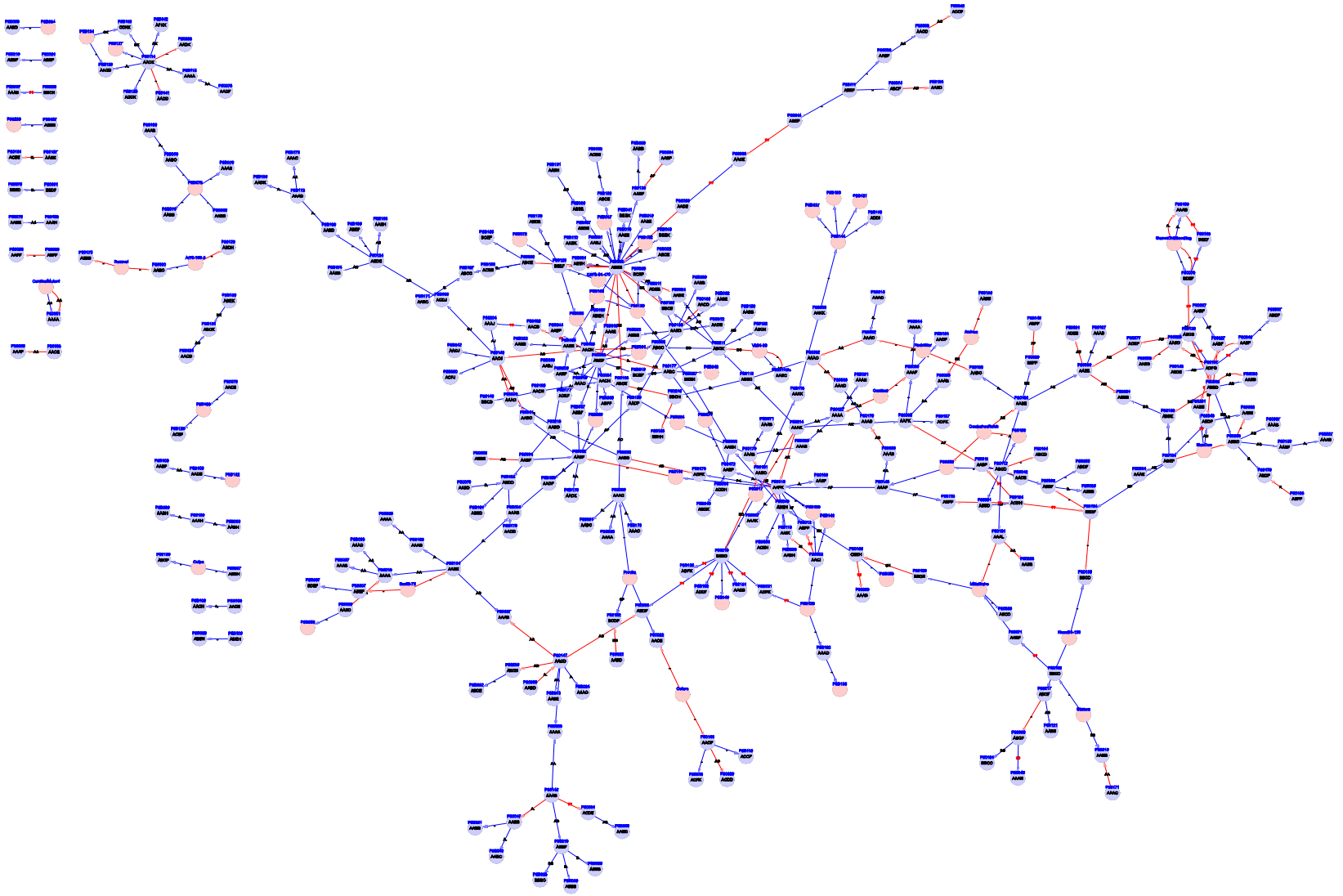

FIGURE S1. Pedigree analysis of tetraploid potato cultivars. The GWD genotype and cultivar code is displayed in the nodes. Red nodes represent cultivars with missing GWD genotype. Arrows point from parent to offspring and are either blue (male parent) or red (female parent).

Genetic diversity

Gallais (2003) proposed the following terms to describe tetraploid genotypes: monogenic (aaaa), digenic-simplex (aaab), digenic-duplex (aabb), trigenic (aabc) and tetragenic (abcd), terms we adopt here. When for a cultivar only one of the GWD amplicons was successfully re-sequenced, the five haplotypes that were identical-in-state in either of the amplicons (A_1, A_2, A_3, A_4 and A_5) could not always be fully resolved. To strengthen the analysis of genotypic variation, these similar haplotypes were grouped into a single haplotype A group. A monogenic condition was observed in nine cultivars that were monogenic for haplotype A. Four of these contained only the major haplotype A_3 and were truly homozygous at the *StGWD* locus. The five other cultivars contained three copies of the A_3 allele and a copy of either the A_1 or A_2 allele. All other cultivars were heterozygous. We observed 77 tetragenic, 185 trigenic, 76 digenic-simplex and 37 digenic-duplex cultivars. The average number of alleles per individual (*Ai*) was 2.86 when the haplotypes A were grouped and estimated at 3.08 when using all 16 haplotypes. A total of 111 different genotypic classes were observed. The number of cultivars per class ranged from 1 to 27 (3.5 cultivars per class on average). The most abundant genotypic class was AAAB occurring in 27 cultivars, followed by AABB, AABD and AABF. Observed and expected heterozygosity (Ho = 0.765, He = 0.758) were in close agreement when assuming random chromosome segregation. A χ^2 test showed that the mean fixation index (F) was in accordance with Hardy-Weinberg expectations.

Associations with starch phosphate content

Starch phosphate content was measured for 203 of the 398 genotyped cultivars. It ranged from 12.6 to 37.7 nmol/mg, with an average of 22.5 ± 4.3. Variation in starch phosphate content within the 80 genotypic classes with measured starch phosphate contents was substantial. Two genotypic classes differed significantly from the other classes. Average starch phosphate content of homozygous class AAAA (14.6 nmol/mg, n=4) was significantly lower than the other classes. The single cultivar representing class BBCH had a significantly higher starch phosphate content (37.7 nmol/mg). Linear mixed model analysis, modeling all haplotypes, identified significant independent associations to starch phosphate content for the grouped haplotype A (p-value 0.009) and haplotype H (p-value 0.015). The haplotype A association explained approximately 13.4% of the populations phenotypic variance and showed a negative association with starch phosphate content. Haplotype H showed a positive association and explained around 4.7% of the variance (Figure 3).

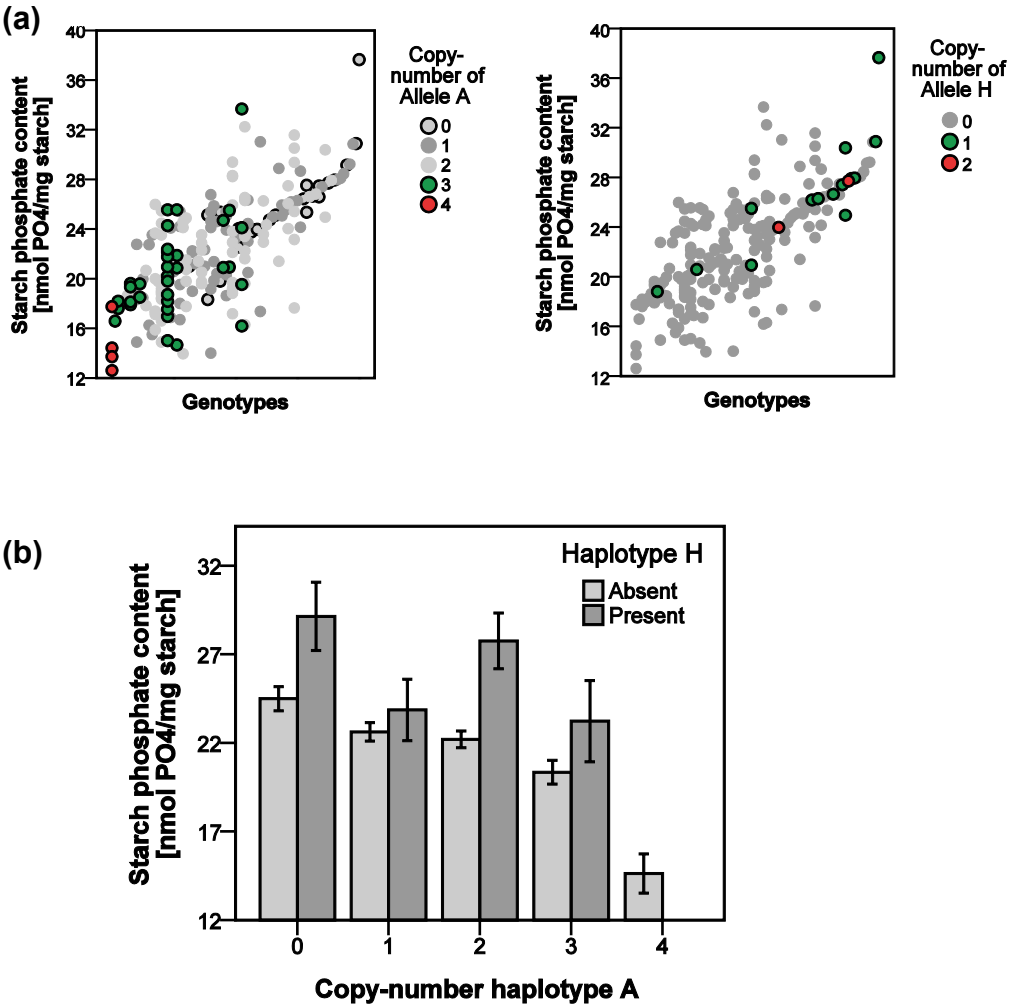

FIGURE 3. Association between GWD allele-copy number and starch phosphate content for haplotypes A and H. (a) Cultivars with the same four-allele genotype (80 different four-allele genotypes were found in 203 cultivars with measured starch phosphate content) have the same x-coordinate. Cultivars with different genotypes are ordered according to the increasing average starch phosphate content of the genotype. **(b)** Combined bar plot of the association between haplotypes A and H and the starch phosphate content of cultivars. Error bars show the standard error of the mean.

Validation in segregating populations

To confirm the association of the haplotype A, the GWD genotypes of 93 plants of the diploid potato C×E mapping population were resolved using HRM. Three distinct GWD haplotypes were observed in the parental genotypes. Haplotype A_2 is shared between both parents, haplotype F is unique to the C-parent and haplotype B unique to the E-parent. Similar to the

results found for the association analysis, the C×E mapping population plants lacking allele A had significant higher starch phosphate content while the offspring homozygous for allele A had significant lower starch phosphate content (Figure 5).

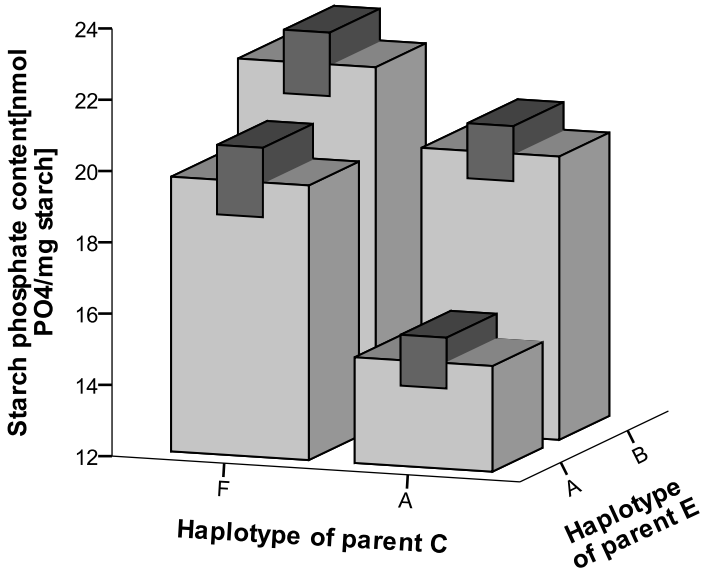

FIGURE 5. Amount of phosphorylated starch in the four genotypic GWD classes of the C×E population. Similar to the results found for the association analysis, the C×E mapping population plants lacking allele A have significantly higher starch phosphate content while the offspring homozygous for allele A have significant lower starch phosphate content. Error bars (dark grey) show the standard error of the mean.

To verify the association of the haplotype A and H we also genotyped 76 tetraploid offspring of a cross Astarte (A_3A_3CI) × Voran (A_3A_3CC) and 34 tetraploid offspring of a selfing population of Sunrise (BBHH) using HMR. Starch phosphate content was measured in 34 offspring of the Astarte × Voran cross and in 19 offspring of Sunrise. For the Sunrise offspring we only obtained offspring genotypes with either no, one or two copies of the H allele. For the Astarte × Voran cross we analyzed only those plants that had allele A_3 and/or C. Offspring with allele H showed a clear tendency (p-value 0.070) towards higher starch phosphate content and offspring homozygous for allele A_3 had a significant (p-value <0.001) lower phosphate content (Figure 4).

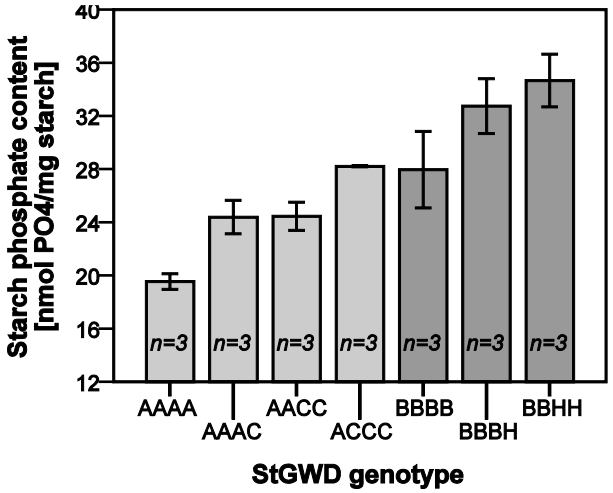

FIGURE 4. Starch phosphate content observed in descendants from a cross between Astarte (AACI) × Voran (AACC) (light grey bars), as well as descendants of selfed Sunrise (BBHH) (dark grey bars). Error bars show the standard error of the mean.

DISCUSSION

Nucleotide diversity

Two regions of the *StGWD* gene of 627 bp and 606 bp were analyzed by direct sequencing of PCR products from monoploid, diploid and tetraploid potato clones. Analysis of the sequence chromatograms, along with verification of a number of haplotypes using cloned PCR products, allowed us to identify a set of 16 haplotypes and their tag SNPs. The tag SNPs were instrumental to fully genotype 384 tetraploid potato cultivars and to study the genetic variation and phenotypic effect of the *StGWD* gene.

DNA sequence variation in potato is exceptionally abundant. We found an overall frequency of polymorphic sites of one variant every 12 bp for the *StGWD* gene. For this large panel of cultivars and progenitor lines, the number of polymorphic sites even exceeds the level of one variant per 21-23 bp found in previous studies (RICKERT *et al.* 2003; SIMKO *et al.* 2006). The frequency of polymorphic sites and the molecular diversity can however vary widely, depending on how many clones, which regions and which genotypes are being analyzed. The study of Simko et al. (2006) involved 47 samples, including some wild accessions, and was re-sequenced for 66 loci. Comparison of nucleotide diversity between Simko et al. (2006) and our study shows that this statistic is more stable across studies ($\pi = 14.6x10^{-3}$ and $\pi = 16.2x10^{-3}$, respectively).

As expected, the average within-potato sequence diversity in our study does not exceed the between species nucleotide diversity at the *StGWD* locus. However, the level of sequence homology of the two most distant potato haplotypes approaches the tomato/potato species

sequence divergence. Considering tomato/potato divergence time of 7.3 MYA (WANG *et al.* 2008), most observed haplotypes seem to have diverged long time ago, at least predating domestication in potato. Some haplotypes were observed only in recent cultivars and likely resulted from efforts to introduce late blight and nematode resistance from related species (BRADSHAW and RAMSAY 2005). Analyzed cultivars harboring haplotype H have a lineage descending from clone USDA 96-56. It is likely that this haplotype has been introduced into the *S. tuberosum* genepool together with the introduction of the chromosome 5 *Phytophthora infestans R1* resistance gene from *S. demissum* (SINDEN and SANFORD 1981). For another haplotype, haplotype A_1 it is likely that it has been introgressed into the cultivated genepool from the CPC 1673 donor of the chromosome 5 potato cyst nematode (*Globodera rostochiensis*) resistance allele *H1* (ELLENBY 1952). And haplotype K seems to be introduced into the genepool by introgression of the chromosome 5 *Globodera pallida* nematode resistance from clone VTN 62-33-3 (ROSS and HUNNIUS 1986).

Pedigree analysis confirms haplotypes are identical-by-descent

The complete haplotype resolution achieved in this study facilitates the study of transmission of alleles from parents to offspring. Offspring-parent pairs should share at least two haplotypes at a locus, since in view of its proximity to the centromere double reduction (leading to two identical haplotypes in a diploid gamete) is exceptional for *StGWD*. Deduced gamete genotypes indeed are in good congruence with those predicted from parental genotype configurations; most offspring genotypes can be explained by their parental genotypes. The analysis of gamete genotypes as inferred from the pedigree data thus confirms the accurate quantification of haplotype copy number and demonstrates that the GWD haplotypes are identical-by-descent. The few inconsistencies between genotyping and pedigree data can be explained by genotyping errors, DNA source errors, or recorded pedigree relationship errors. To rule out genotyping errors we re-analyzed the sequence chromatograms for other than the selected tag SNPs, in both GWD amplicons, and confirmed the called genotypes are consistent. For four genotypes, inconsistent in multiple parent/offspring cases, it is most likely that the DNA source we used for sequencing does not correspond to the genuine cultivars. In the 10 remaining cases either the DNA source of the offspring, the involved parents or the recorded pedigree relationships are likely to have caused the mismatch. Since there are only few cases (10/218 <5%) were the pedigree relationship data do not coincide with the genotype data and in these cases mistakes in DNA source cannot be ruled out, this study demonstrates the high accuracy of the potato pedigree database (VAN BERLOO *et al.* 2007).

Genetic diversity

The haplotype tag SNPs, with a quantitative scoring of allele copy number for two independent amplicons, gave us the possibility to exploit the full genotypic information. We used this genotypic information to evaluate the genetic diversity in the analyzed set of cultivars and accessions. Intra-individual heterozygosity (Ho = 0.77) and the mean observed number of haplotypes per plant ($Ai \approx 3.1$) are high at the *StGWD* locus. Both the number of haplotypes and heterozygosity are markedly higher than those reported in an earlier allozyme

study of 13 loci in tetraploid potato cultivars (OLIVER and MARTÍNEZ ZAPATER 1984), which seems to demonstrate the superior resolution of SNP markers compared to allozyme studies. The number of observed haplotypes was also higher than those reported in more recent potato re-sequencing studies (LI *et al.* 2005; SATTARZADEH *et al.* 2006; SIMKO *et al.* 2004) and comparable to a multi-locus study employing SSR markers on the same cultivar set (D'HOOP *et al.* 2010). Furthermore, the above average levels of alleles and heterozygosity observed in individual potato cultivars suggest that an underestimation of heterozygosity caused by allele homoplasy should be of a minor magnitude and strengthen our conclusion that the alleles are identical-by-descent. Full allelic resolution of the *StGWD* locus would however require the complete gene to be re-sequenced, while in this study we only sequenced parts of the gene. Therefore, it cannot be excluded that currently unresolved haplotypes, identical-in-state to the identified haplotypes, remain in the genepool. Near complete re-sequencing of the *StGWD* gene in 84 potato cultivars and accessions using massively parallel sequencing has however not identified new alleles of the *StGWD* gene (see Chapter 4).

Association analysis of StGWD haplotypes with starch phosphate content

In potato there is only a small diversification into subpopulations. This diversification is along cultivars used for fresh consumption, processing (chips, fries), and potatoes used for the starch industry (D'HOOP *et al.* 2010). We did not observe differences in *StGWD* allele frequencies in these subpopulations and therefore omitted correction for population structure in the association analysis.

Starch phosphate content is hardly influenced by environmental conditions (HAASE and PLATE 1996; WERIJ *et al.*) and can be measured with a small technical error (NODA *et al.* 2006). A large variation in starch phosphate content within each genotypic class of GWD alleles was however observed in the tetraploid association mapping panel. This large variation indicates that, similar to the diploid C×E mapping population, in tetraploid cultivars multiple QTL with major effects on different genomic locations are associated with the trait.

Starch phosphate measurements and QTL analysis in the diploid C×E potato mapping population has been described previously (WERIJ *et al.*). The QTL analysis showed three major additive QTLs on chromosomes 2, 5 and 9, each explaining approximately 20% of the observed variance. The QTL on chromosome 5 co-localized with the *StGWD* locus, a key enzymes involved in starch phosphorylation (ZEEMAN *et al.* 2007). We re-sequenced the parental genotypes of the C×E population and identified the GWD haplotypes in the C×E offspring using HRM. The reducing effect of haplotype A on starch phosphate content detected by the association analysis, is confirmed in the C×E population. Additionally, we verified the phenotypic effect of haplotype A and haplotype H in two tetraploid cross population. Results from the present study thus indicate that a haplotype association analysis approach is a robust tool for mapping quantitative loci with relatively strong effects in commercially important potato populations, even without considering population structure.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Nick de Vetten of Averis Seeds, Valthermond, the Netherlands, for producing and phenotyping the two validation crosses. We thank Heleen Furrer for measuring starch phosphate content. This research was supported by a grant of the Dutch technology foundation STW, project WPB-7926.

CHAPTER 4

A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato

AUTHORS
J.G.A.M.L. Uitdewilligen
A.M.A. Wolters
T.J.A. Borm
H.J. van Eck
R.G.F. Visser

ABSTRACT

Genotype calling in autotetraploids implies the ability to distinguish among five possible alternative allele copy number states. This study demonstrates the accuracy of genotyping-by-sequencing (GBS) in a large collection of autotetraploid potato cultivars using next-generation sequencing. Sufficient read depth and population size can be obtained, within acceptable costs, when using genome complexity reduction. We enriched cultivar-specific sequencing libraries for 2.1 Mb of the potato genome, covering ~800 distributed target genes, with a custom-designed SureSelect hybridisation in-solution enrichment library. Indexed sequencing libraries of 83 tetraploid cultivars and one reference accession were paired-end sequenced in 7 pools of 12 samples using the Illumina HiSeq2000. After filtering and processing the raw sequence data, 12.4 Gigabases of mapped, high-quality sequence data was obtained, with a median average read depth of 63× per cultivar. We detected over 129,000 sequence variants and estimated the minor allele frequencies and zygosity type frequencies of tetraploid samples in the potato population. The average nucleotide diversity varied among the twelve potato chromosomes and individual genes under selection were identified. Validation of sequence-based zygosity estimates with a SNP genotyping assay indicated the accuracy of GBS for autotetraploids at a read depth of 80×. The genotype data were tested in a marker-trait association study, and allowed the identification of common alleles strongly influencing plant maturity and potato tuber flesh colour. As an alternative to GBS, we present a high-throughput whole-genome SNP array based on the sequence variants identified in this study.

INTRODUCTION

DNA variants such as single nucleotide polymorphisms (SNPs), multinucleotide polymorphisms (MNPs), and insertions and deletions (indels) are differences at the nucleotide sequence level among individuals or alleles and represent the basic units of genetic diversity. They can be assayed and exploited as high-throughput molecular markers and are widely used for marker assisted and genomic selection, association analysis and mapping of quantitative trait loci (QTL), and haplotype and pedigree analysis in several crops and model plants (ATWELL *et al.* 2010; CLOSE *et al.* 2009; GRATTAPAGLIA *et al.* 2011; HAMILTON *et al.* 2011; HUANG *et al.* 2010; JANNINK *et al.* 2010; MOOSE and MUMM 2008). Genotyping of DNA sequence variants in highly heterozygous polyploid species, such as potato (*Solanum tuberosum*), is more challenging than in diploid species, because a given gene may be represented by up to four different alleles per locus per genotype. Therefore, genetic analysis in tetraploid potato requires a genotyping system that can distinguish among alleles and quantify the allele copy number. In tetraploid species, the possible allele copy number (zygosity) categories include nulliplex (0), simplex (1), duplex (2), triplex (3), and quadruplex (4).

Several studies have shown that direct resequencing of amplicons by Sanger sequencing is sufficiently quantitative to allow the simultaneous discovery and genotyping of sequence variants in polyploids (RICKERT *et al.* 2002; SATTARZADEH *et al.* 2006). Amplicon sequencing is a reasonable method to analyse a single or small number of target gene(s), but beyond that

scale, more effort is required to design unique primers and optimize PCR parameters to ensure equal amplification of all alleles. Furthermore, in species like potato that exhibit high nucleotide diversity, the many indels usually create uninterpretable sequence reads that further reduces the throughput of Sanger-based amplicon resequencing. With the introduction of the next-generation massively-parallel sequencing (MPS) technologies, several novel approaches have been developed to discover, sequence and genotype not tens, but hundreds to thousands of genes simultaneously (CRAIG *et al.* 2008; GORE *et al.* 2009; GRIFFIN *et al.* 2011; HAMILTON *et al.* 2011; METZKER 2005; MYLES *et al.* 2010). Similar to Sanger sequencing, MPS can be used directly to genotype sequence variants. Using the former technology, genotype data are retrieved from chromatogram peak intensities at variant positions. For MPS, accumulation of sequence reads provides digital allele estimates for each variant. Accuracy of this allele estimate is dependent on sequencing depth. In many studies, whole-genome resequencing utilizes low read depth (e.g., <5× per site per individual); for diploids, the probability that only one of the two chromosomes has been sampled at a specified site is relatively high. Binominal distribution and a simulation model (GRIFFIN *et al.* 2011) have suggested that ensuring all four homologous chromosomes of a tetraploid organism are sequenced at least once requires a read depth of approximately 15×. From a binominal distribution, we estimate that the probability of a correct duplex call requires a sequence depth of at least 48× (See Chapter Introduction).

To achieve an increase in read depth, the portion of the genome that is sequenced can be reduced by applying, for example, RNAseq (GORE *et al.* 2007; HAMILTON *et al.* 2011), restriction enzyme based complexity reduction (ELSHIRE *et al.* 2011; VAN ORSOUW *et al.* 2007), or sequence capture methods such as SureSelect, Nimblegen, and Raindance (GNIRKE *et al.* 2009; KISS *et al.* 2008; NIJMAN *et al.* 2010). RNAseq is not suitable for genotyping-by-sequencing (GBS), since alleles that vary in transcription level will generate an inaccurate genomic representation of alleles in an individual. Methods based on restriction enzyme treatments, on the other hand, are more likely to target non-coding parts of the genome, producing less useful data for functional gene analysis. Furthermore, restriction-based methods cannot target specific regions of interest, and nucleotide variants in the restriction site may interfere with digestion and cause null alleles . Sequence capture methods like SureSelect use oligonucleotide baits designed to bind to regions of interest, which can be specifically selected and enriched before sequencing (DAVEY *et al.* 2011). For example, whole exomes (MAMANOVA *et al.* 2010) or regions associated with particular traits (GNIRKE *et al.* 2009; HODGES *et al.* 2009) can be targeted. Sequence capture approaches require *a priori* availability of sequence data to design DNA capture probes. The recently sequenced genome of *Solanum tuberosum* group Phureja is appropriate for this purpose in potato (XU *et al.* 2011). When a high-quality, reference sequence from an organism closely related to the study population is available, these methods are potentially highly accurate (GNIRKE *et al.* 2009). This may be less true when the population has considerably diverged from the reference or exhibits high levels of diversity; the sequence-specificity of reference-designed baits leads to weaker hybridization with targets strongly diverged from the reference. This may bias the complexity reduction step against highly polymorphic regions. An advantage of in-solution hybridization, however, is that the long capture probes used are more tolerant to polymorphisms than the shorter sequences typically

used, for example, as primers for PCR or multiplex amplification (GNIRKE *et al.* 2009). Densely tiled capture probes can also increase the likelihood of a bait binding to all alleles of target regions.

In this paper, we describe DNA resequencing results obtained from dozens of autotetraploid potato cultivars and one monoploid accession after genome complexity reduction using hybridisation-based in-solution enrichment. We subsequently used this data to identify sequence variants within and across cultivars and call the genotypes of resequenced individuals. The accuracy of genotyping-by-sequencing was validated using a SNP genotyping assay. The resulting marker dataset is useful for describing allele frequencies, nucleotide diversity, and population structure in potato, and for validating QTLs via association analysis. Our approach is an efficient means of producing data for the design of both high and low-density SNP genotyping assays applicable to a wide range of potato cultivars, and the resulting tools can be used to address questions in population genetics and marker-trait association research.

MATERIALS AND METHODS

Design of the capture library

A custom SureSelect capture library containing >57,000 RNA oligonucleotide baits of 120-bp length each was designed using seven cDNA sequence databases (Table 1). Genes targeted for enrichment include subsets of those in the PotCyc metabolic pathway database (MENDA *et al.* 2008), the Potato Maps and More database (PoMaMo; MEYER *et al.* 2005), an in-house QualitySNP marker database (ANITHAKUMARI *et al.* 2010; TANG *et al.* 2006), and a subset of single-copy genes homologous to the Conserved Ortholog Set II (COSII; WU *et al.* 2006). Functional genes widely used as genetic markers for carbohydrate metabolism (GEBHARDT *et al.* 2001) and secondary metabolism (MARTIN *et al.* 2004; MCCUE *et al.* 2007; WOLTERS *et al.* 2010), and a number of putative candidate genes for potato quality traits, were also included as targets. In addition to these functional genes, a number of intergenic sequences corresponding to AFLP markers, and a number of chloroplast genes and mitochondrial genes were included (Table 1). For most targets, cDNA sequences were aligned to the *S. tuberosum* Group Phureja DM whole genome assembly v.1 and baits were designed based on the most highly homologous genomic reference sequence. Genome annotation for the DM sequence was not yet available at the time the baits were designed, so genomic coding regions and intron/exon boundaries were estimated using GeneSeqer (USUKA *et al.* 2000). Other reference sequences used for bait design included the chloroplast genome of *Solanum tuberosum* cv. Desiree (NC_008096), *Solanum tuberosum* mitochondrial sequences (S66866, X74826, X80386, X83206, X93576) and a few sequenced BAC clones of genotype RH89-039-16 (XU *et al.* 2011).

During bait design, we aimed to optimize the in-solution hybridisation enrichment by avoiding targets exhibiting repeat elements and paralogous sequences in the reference genome, which can affect target-bait hybridization during enrichment and add difficulty to read mapping after resequencing. Stretches of repetitive sequences within the target regions

were excluded from bait design using RepeatMasker (http://www.repeatmasker.org). BLAST homology search against the *S. tuberosum* Group Phureja DM whole genome assembly v.1 (XU *et al.* 2011) was conducted to avoid the use of targets with paralogs and/or duplications. Except for a small number of target gene families of specific interest (e.g. polyphenol oxidases), target sequences having a secondary hit with E-value $<10^{-10}$ were excluded from bait design. For each target sequence, we used an average of 3-4 continuous regions for bait design, each region having an average length of 475 bp and, where applicable, including both exons and introns. Regions consisting mainly of introns (>200-1000 bp) were avoided as targets. OligoTiler (BERTONE *et al.* 2006) was used to tile the reference strand of each continuous target region, with baits (of 120 bp in length) starting approximately every 20 bp. This produced a 6× bait tiling coverage and resulted in 57,054 unique baits for the SureSelect capture library (ELID 0274451). In total, the library targets 2,945 uninterrupted regions (1.44 Mb, GC-content 39%; Table 1). Complete lists of sequencing targets and oligonucleotide bait sequences are available in XLS-file S1 and FASTA-files S1&S2.

TABLE 1. Databases used for SureSelect capture potato library development. For most targets, database (cDNA) sequences were aligned to the potato reference genome, and the best homologue in each case was used to design the bait.

Database	*Genes*	*Target Contig Regions*	*Target Sequences*	*# of Baits (1 bait / 20 bp)*	*Proportion of baits in library*
COSII database	248	853	439.6 kb	17,618	30.9%
PotCyc database	149	467	142.5 kb	4,940	8.7%
PoMaMo and Candidate genes	116	523	283.1 kb	11,481	20.1%
In-house database	249	789	424.7 kb	17,174	30.1%
AFLP sequences	45	202	87.8 kb	3,400	6.0%
Chloroplast	64	99	55.9 kb	2,272	4.0%
Mitochondrial	4	12	4.7 kb	169	0.3%
Total	**875**	**2,945**	**1,438 kb**	**57,054**	**100%**

Plant collection

A subset of 83 tetraploid potato cultivars (Table S1) was selected using AFLP-based genetic distance estimates from a larger collection described by D'hoop et al. (2008). The subset represents a diverse range of commercial potato cultivars with respect to country of origin, year of release, and market segment. We also included a monoploid potato clone 1-3 511, derived by anther culture of the heterozygous diploid *S. tuberosum* group Phureja, clone BARD 1-3 of accession PI225669 (VEILLEUX and LIGHTBOURN 2007). The 1-3 511 clone is highly related to the recently sequenced clone DM 1-3 516R44 (DM, CIP801092; XU *et al.* 2011), a doubled monoploid derived from the same BARD 1-3 clone (VEILLEUX and LIGHTBOURN 2007).

Extraction and fragmentation of DNA

DNA was extracted from leaves ground in liquid nitrogen using KingFisher Genomic DNA Purification Kit (Thermo Scientific) and the KingFisher Ml magnetic nucleic acid extraction system (Thermo Scientific) according to the manufacturer's procedures. DNA concentrations were quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific) and diluted to 35 ng/µl. Of each DNA sample, 3.5 µg was fragmented by Adaptive Focused

Acoustics on a Covaris S2 instrument (Covaris, Inc.), using a 10% duty cycle at intensity 4 for 120 seconds with 200 cycles per burst. Fragmentation of DNA to an average size of about 300 bp was verified using Bioanalyzer High Sensitivity DNA chips (Agilent).

Indexed sequence library preparation

A paired-end sequencing library using twelve different custom-indexed adapter pairs was prepared to allow the identification of individual potato genotypes from sequence reads taken from pooled DNA samples. The indexed adapters consisted of complementary Illumina adapters PE1 and PE2 (CRONN *et al.* 2008), with PE1 extended by a 4-bp index sequence and an extra terminal T to facilitate sticky-end ligation. The reverse (PE2) adapter was extended by the reverse complement of the PE1 index (Table 2). The twelve indices had a balanced base composition and a minimal edit distance (i.e. the number of mutations required to change one index to another) of 2 bp to detect sequencing errors in the index region.

TABLE 2. Sequences of the twelve indexed adapter pairs. Nucleotides in bold are included to facilitate sticky-end ligation. The index sequence is underlined. Oligos were HPLC purified.

Adapter pair	*Strand A (5' –> 3')*	*Strand B (5 '–> 3')*
PEM01	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCA***T**	P-TGCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM02	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAG***T**	P-CTGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM03	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTACG***T**	P-CGTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM04	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGC***T**	P-GCACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM05	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTC***T**	P-GACCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM06	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCGT***T**	P-ACGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM07	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGA***T**	P-TCAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM08	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTA***T**	P-TACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM09	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCATG***T**	P-CATGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM10	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATAC***T**	P-GTATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM11	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAT***T**	P-ATGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
PEM12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACT***T**	P-AGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

* = phosphorothioate bond
P- = phosphate group

To pair the adapter strands, mixtures of the forward and reverse strands (each 50 μM in TE) were incubated at 95 °C for 2.5 minutes, followed by a cool down (-1 °C/30 sec.) to 25 °C and subsequent dilution to a working concentration of 12.5 μM. DNA sequencing libraries were prepared using the NEBNext DNA Sample Master Mix Set 1 (New England Biolabs). Purifications were carried out between end-repair, dA-tailing, and adapter ligation steps using AMPure XP beads (Agencourt Bioscience). To tag the 84 potato DNA extracts, each of the twelve unique adapters was ligated to seven different DNA samples by mixing adapter and DNA in a molar ratio of approximately 20:1. After the initial ligation of 15 minutes at 20°C, samples were held at 4°C overnight and purified using AMPure XP beads. Adapter ligation was verified by fragment size analysis using Bioanalyzer High Sensitivity DNA chips.

In-solution hybridization and target enrichment

The 84 indexed paired-end sequencing libraries were hybridized to the SureSelect capture library according to the manufacturer's instructions (Agilent SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library Protocol, Version 1.0 May 2010), with minor modifications. Size selection on gel was omitted, allowing us to skip the standard pre-amplification step prior to in-solution enrichment. This reduces the number of clonal reads in the later generated sequences. The use of index-specific blocking oligos to reduce non-specific pull-down due to adapter-adapter hybridisation can also be avoided when pre-amplification is excluded, as non-amplified Y-adapters do not concatenate. We used half of the specified hybridization volumes, and excluded from the hybridization mix Block #1 (Human C0t-1 fraction; HARISMENDY *et al.* 2009), which is irrelevant for plants), and Block #3 (PE adapter block; HARISMENDY *et al.* 2009). Of each indexed DNA sample, 50-400 ng was mixed with 1.25 μg salmon sperm DNA (Block #2), denatured, and hybridized to 100 ng SureSelect biotinylated RNA baits developed from the capture library.The hybridization mix was held at 65°C for 24 h for hybridization, then added to 250 ng (25 μl) T1 streptavidin Dynabeads (Invitrogen), and pulled down. Bait-selected DNA was purified using AMPure beads, and amplified for 14 cycles (Tm 60 °C) using 1 U Herculase II Fusion proofreading DNA polymerase (Agilent) and 25 μM each of custom primers PE1.0 and PE2.0 (Table 3). The alternative DNA polymerase was chosen after the polymerase included in the NEBNext DNA Sample Master Mix Set 1 was seen to cause a negative shift in average fragment size.

TABLE 3. Primers used for amplification of hybrid-selected DNA.

Name	*Sequence (5' –> 3')*
PCR_PE2.0	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T
PCR_PE1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T

* = phosphorothioate bond

Amplified, enriched DNA libraries were purified using AMPure XP beads and evaluated by agarose gel electrophoresis, NanoDrop, and Bioanalyzer High Sensitivity DNA chips. Library subsamples were also quantified by qPCR using a two-step amplification protocol [95 °C activation for 2 min, 40× (95 °C 10 s, 60 °C 30 s)] and primers qPCR_1.1 and qPCR_2.1 (Table 4).

TABLE 4. Primers for quantification of enrichment libraries.

Name	*Sequence (5' –> 3')*
qPCR_1.1	AATGATACGGCGACCACCGAGAT
qPCR_2.1	CAAGCAGAAGACGGCATACGA

Hi-Seq2000 sequencing and data preprocessing

Seven distinct resequencing pools were made by combining equimolar amounts of twelve unique, differently-indexed, enriched DNA library samples (Table S1). Pooled libraries were sequenced on a Hi-Seq2000 (Illumina) lane using 100-base paired-end sequencing at the Genome Analysis Facility of the University Medical Center Groningen (UMCG). Four sequenced pools that generated low cluster numbers and/or partly failed reverse read sequences were repeated. The sequence data analysis pipeline is summarised in Figure 1. Initial quality checks (average read quality per cycle, average read quality, base call % per position) were performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequences in each pool were categorized to their biological sources according to the 4-bp index using NovoBarCode (Novocraft) with the average edit distance set to 3. The first 5 bp of each de-multiplexed sequence read (i.e., the 4-bp index plus an extra T used for ligation) was removed, and a code representing the potato genotype was added to the read name.

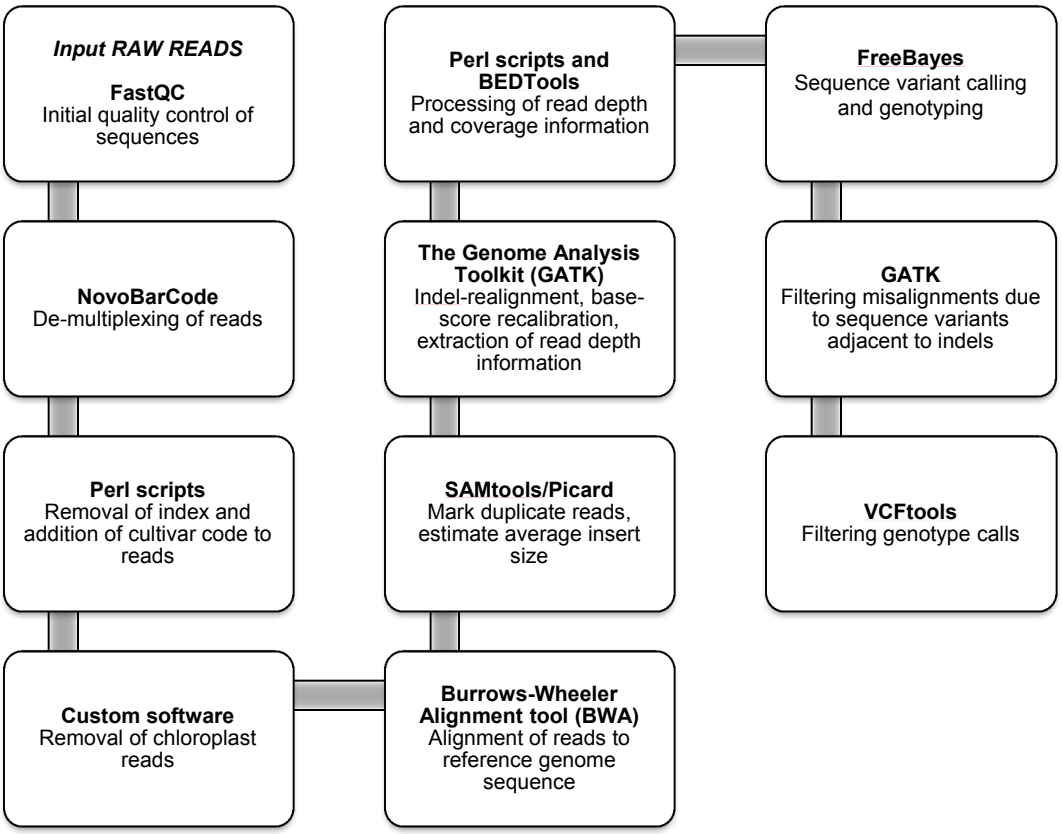

FIGURE 1. Overview of the sequence data analysis.

Highly abundant sequences containing ≥63 continuous nucleotides with 100% match to chloroplast genome sequences of *Solanum tuberosum* cv. Desiree were extracted from the de-multiplexed sequences and kept for separate analysis. Remaining sequences were aligned using the Burrows-Wheeler Alignment tool (BWA; DURBIN and LI 2009) with default alignment parameters, except for the maximum edit distance, which was relaxed to seven due to the expected high sequence divergence between potato alleles. Sequences were aligned to the annotated superscaffolds (DMGv.3.4, comprising 705.5 Mb of sequence in 1,419 superscaffolds and 43 Mb of non-ACGT bases) of the *S. tuberosum* Group Phureja DM whole genome assembly v.3 reference genome (XU *et al.* 2011). The 120-bp SureSelect bait sequences of genomic origin were correspondingly mapped to the annotated superscaffolds, and uniquely mapped baits with a mapping quality (MQ) ≥37 were used to define the genomic target regions.

Alignment data was processed with SAMtools and Picard (DURBIN *et al.* 2009) to mark duplicate reads and estimate the average insert size of the paired-end reads, then read-group and sample information was added to the aligned sequences using Perl scripts. The Genome

Analysis Toolkit (GATK; MCKENNA *et al.* 2010) was used for indel realignment, base-score recalibration, and extraction of read depths. Read depth and coverage data were processed with Perl scripts and BEDTools (QUINLAN and HALL 2010).

Sequence variant detection and genotype calling

For covered regions (see Results for definition), sequence variants were identified simultaneously among the aligned reads from all 83 tetraploids and the single monoploid using the FreeBayes polymorphism discovery algorithm (The MarthLab). Sequence variants included binary SNPs, MNPs, and small indels, as well as allelic series of tri-SNPs and tetra-SNPs, multi-allelic MNPs, and indels with a variable number of (repetitive) nucleotides. Reads with more than seven base mismatches, more than three separate gaps, or with MQ<30 were excluded from variant calling. The expected mutation rate or pairwise nucleotide diversity was set to 0.01. In order to include an alternate allele as a variant, supporting bases required a minimum base quality (BQ) of at least 13, and at least one supporting alignment was required to have BQ≥20. Furthermore, the alternate allele had to be observed in at least 5 reads and represent at least 12.5% of the observations of reads at that locus within a single potato sample. Mapping quality of alleles was included when calculating posterior probabilities, and variants were only called for sites that had a probability of polymorphism greater than 0.95. Sequence variants adjacent to indels, which were likely due to local misalignment, were filtered using GATK.

Zygosities of all sequence variants were resolved by allele-specific read depths for all reads with MQ≥13 using FreeBayes. This resulted in nulliplex (0/0/0/0), simplex (1/0/0/0), duplex (1/1/0/0), triplex (1/1/1/0) and quadruplex (1/1/1/1) genotype calls relative to the reference sequence for the tetraploid potato samples, while the monoploid sample was genotyped as either absent (0) or present (1) for each variant. Only variants previously identified by variant calling were used for genotyping, and samples required a minimum read depth of 15× at the variant position to yield a genotype call.

Diversity analysis and prediction of functional consequences of allelic variants

To calculate the nucleotide diversity, we first calculated the gene diversity (heterozygosity) of each binary variant; $GD = 1 - (p_{minor}^2 + p_{major}^2)$, where p_{minor} is the frequency of the minor allele and p_{major} the frequency of the major allele. GD is regularly used to determine the value of a marker in detecting polymorphism, i.e. the polymorphisms information content (PIC), and can be extended to multi-allelic markers. In such a case, $GD = 1 - \Sigma P_i^2$, where P_i is the frequency of the *i*th allele and GD is summed across *n* alleles. Nucleotide diversity (π) was calculated by averaging GD over all nucleotides sited – or all coding and non-coding nucleotides sited – on a covered fragment (contig) or gene.

The functional effect (codon mutations, splice site mutations, frame shift mutations) of genetic variants was predicted using snpEff (CINGOLANI *et al.* 2012) and gene annotation version DMGv3.4 of the DM genome.

In order to identify population structure, principal components analysis (PCA) was performed on genotype scores using the FactoMineR library of R (Lê *et al.* 2008). Only genotype calls of variants found in all 84 cultivars were included. Since all genotype scores were measured in units of allele copy number, the data were not scaled. K-mean clustering of the first three principal components was used to identify genotype clusters.

Association analysis

Adjusted phenotype means for plant maturity and tuber flesh color in each of the 83 tetraploid cultivars, measured previously over a period of five years (D'HOOP *et al.* 2010; D'HOOP *et al.* 2008), were used as trait values for conducting association analysis. Additive and dominant genotype models were each tested with and without correction for population structure. The genotype clusters identified by PCA analysis were used as the adjustment factors for population structure. For dominant association models, linear regression models were implemented in PLINK (PURCELL *et al.* 2007) using only binary allelic variants and applying the permutation procedure to generate empirical significance levels. Adjustment of p-values to correct for multiple testing effects was carried out using step-up FDR control. For additive genotype models, we applied linear regression models implemented in Genstat. For each trait and each marker the model fitted was: response = allele copy number (+ structure) + error.

Validation of genotype calling

The accuracy of GBS genotype calls was validated using the Kbioscience Allele-specific Polymorphism Assay (KASP) SNP genotyping platform. Binary SNPs identified in our resequencing data that exhibited a minor allele frequency between 0.15 and 0.35 were selected as candidate for assay design. To assure independence among all SNPs in the final design (i.e., use of a unique SNP from each haplotype block), SNP data from GBS were clustered using hierarchical cluster-analysis, and a single SNP from each cluster having a correlation coefficient of $r^2 \geq 0.16$ was used for the KASP assay, yielding 768 SNPs in the final design. We assayed DNA from 65 potato cultivars included in GBS, with two of these cultivars measured in duplicate to assess KASP genotyping consistency. A number of additional diploid potato clones were assayed (~96) and used to examine the cluster position of the nulliplex, duplex and quadruplex genotype signals. The software package fitTETRA (VOORRIPS *et al.* 2011) was used for full tetraploid zygosity genotype calling. In total, 270 of the 768 KASP assayed SNPs were selected for validation of the GBS calls. This selection was based on (1) sharp clustering of the signal ratios of discrete genotype classes, (2) clustering signal ratios of heterozygous diploids with duplex tetraploids, and (3) clustering of signal ratios of homozygous diploids with nulliplex or quadruplex tetraploids. For one of the cultivars duplicated in the KASP assay, all 270 genotype calls were identical between replicates, and for the other cultivar, only two genotype calls varied between replicates. The genotyping error of this subset of 270 selected SNPs is thus very low (0.4%). For concordance analysis, expected genotype calls were obtained from the 270 KASP assay SNPs and observed genotype calls were obtained by the GBS results.

RESULTS

Sequencing and mapping of enriched targets from 84 potato genomes

We designed an in-solution hybridisation capture library targeting primarily introns and exons of nuclear coding genes, but also including some intergenic, chloroplast, and mitochondrial sequences. Baits targeting nuclear genome sequences in the enrichment library covered approximately 1.3 Mb of the potato DM reference genome sequence, scattered across all 12 chromosomes (Figure 2 & BED-file S1).

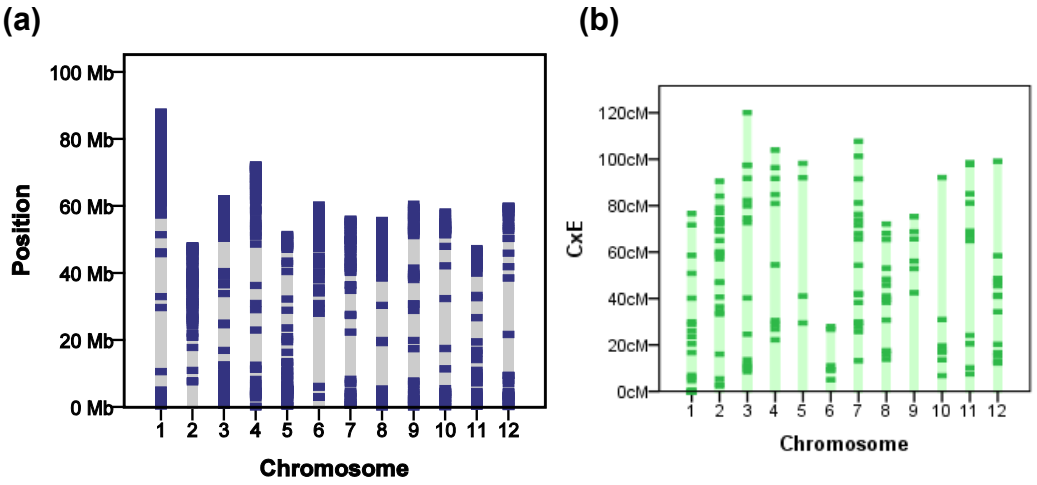

FIGURE 2. Mapping of resequencing targets onto the potato genome. Panel A**:** Physical position of the approximately 800 target genes scattered across the 12 potato chromosomes. Superscaffolds are ordered according to the latest pseudomolecule order (v.3 beta) and show complete chromosomes according to current knowledge. Intergenic targets are not shown. Panel B**:** Genetic position of 250 of the target genes, mapped in C×E. Segments of some chromosomes are not shown, and the two panels do not show chromosomes in the same alignment.

Genomic DNA sequencing libraries from 83 genetically diverse tetraploid potato cultivars and progenitor lines, and a single monoploid potato clone, were indexed by cultivar with 12 distinct 4-bp index sequences and individually enriched using the capture library. Indexed and enriched samples were multiplexed in pools of 12 unique indices and paired-end sequenced. In total, 592,100,112 read-pairs were obtained, representing approximately 100 Gigabases of sequencing data. The cultivar-specific sequence index could be identified for 96% of read-pairs (Figure 3). For 19% of the read-pairs, cultivar identification was based on the index from only one read of the pair. In most cases, this resulted from technical failure of the reverse-read sequencing.

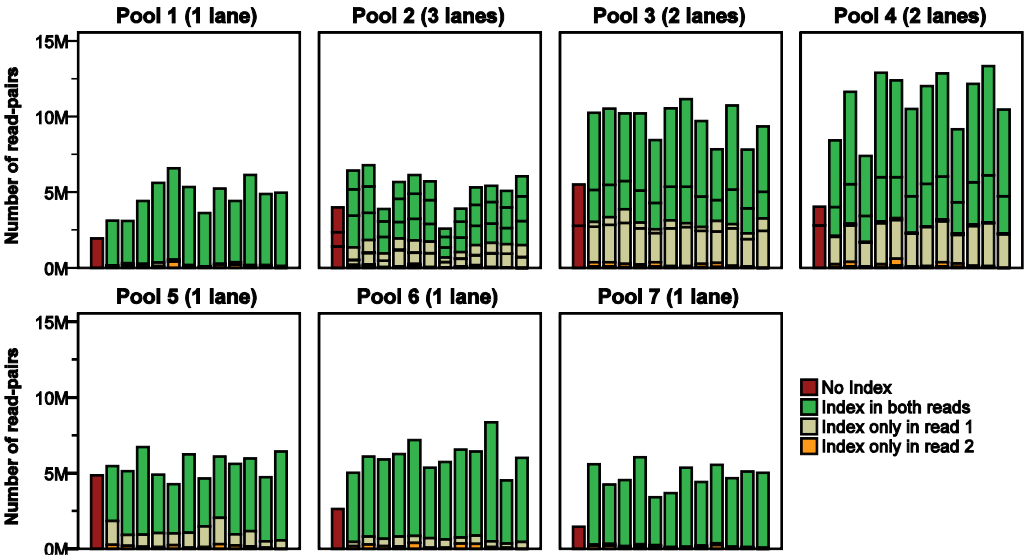

FIGURE 3. HiSeq2000 read-pairs. Each graph represents the number of read-pairs obtained from twelve potato cultivars, loaded as a single pool, for a total of seven pools. Most pools required only one sequencing lane, but pools 2, 3, and 4 exhibited reverse-read sequencing failures and therefore required multiple sequencing lanes, as noted in their respective panels. In total 11 HiSeq2000 lanes were sequenced. Green bars represent the number of read-pairs with valid indices in both reads. Yellow, orange and red bars represent the number of read-pairs where the forward, the reverse, or neither of the indices were valid. Coloured bars interrupted with a line indicate separate read-pairs obtained during the first attempt (lower part) or from the second or third attempts (middle or upper part of the stacked bar, respectively).

Within each multiplexed pool, the cultivar-specific read counts rarely reached a twofold difference between cultivars, indicating approximately equivalent amounts of DNA were added to a pool from each cultivar present. Initial inspection of the cultivar-tagged sequences showed that chloroplast-derived sequences were highly abundant (60% of the read-pairs). These sequences were filtered using broad homology criteria and saved for a later analysis of its own. The remaining 227,263,706 read-pairs were aligned to the DM reference genome, with 80% mapping. This resulted in 23.9 Gigabases of high quality (MQ≥13) genome-aligned sequence data. Among mapped reads, 8% were marked as duplicate reads aligning with identical start and end positions on the reference genome.

Genome and target coverage

As a consequence of the enrichment method, which allows the possibility of genomic DNA partially hybridizing to a bait and including additional unbound, non-targeted DNA in the enriched sample, some sequences aligned not only to target regions but also to flanking and off-target regions. To define the "accessible" part of the genome, i.e., the part with adequate coverage to allow identification of sequence variants, we used the following criteria: (1) at least five cultivars had to be sequenced at the base position, with a minimal read depth of 20× for very high mapping quality reads (MQ≥30) per cultivar; (2) in view of the fragment sizes of

mapped reads (261±94 bp, mean±SD), regions defined as covered were at least 261 bp long; (3) regions were discarded when more than 5% of reads within the region aligned to multiple locations with equal probability (MQ=0); and (4) finally, regions with homology to chloroplast and mitochondrial sequences were removed. This resulted in 2,445 uninterrupted covered regions (contigs) with a total length of 2.1 Mb (XLS-file S2 and BED-file S2). A total of 12.5 Gigabases of high quality sequence data aligned to these contigs, providing an average read depth of 5,871× when all cultivars were considered together. Median depth per cultivar within the accessible genome was 63×, ranging from 15× to 177× (Table 5 and Table S1).

TABLE 5. Summary of target enrichment sequence coverage for 84 potato cultivars.

"Accessible" genome parameter	*Value*
Number of covered regions (contigs)	2,445
Genes covered	977
Sequence length	2,136,143 bp
Coding sequence (DMG v.3.4)	655,930 bp
Target genes covered	793 out of 807
Target sequence covered	1,294,097 bp (97%)[a]
Target + directly flanking sequences	1,848,192 bp
Off-target sequence	287,951 bp
Mapped sequences per cultivar [b,c]	149 ± 76 Mb
Sequencing depth per cultivar [c]	70 ± 36× (Median 63×)
Nucleotide diversity (×10-3) [c]	10.7 ± 5.4 SD
Nucleotide diversity (×10-3), coding sequence [c]	7.3 ± 4.8 SD
Nucleotide diversity (×10-3), non-coding sequence [c]	12.4 ± 6.7 SD

[a] Percentage of target bait sequence mapped to the DM genome.
[b] The sum of sequenced nucleotides aligned with high quality (MQ≥13).
[c] Mean and standard deviation.

Almost all genomic regions and genes targeted by the enrichment library fell within the accessible regions, with 97% of target sequences and 793 of 807 genomic target genes covered. In total, 10.7 Gigabases of high quality sequence data aligned to target regions with a median depth per cultivar of 88×, ranging from 20× to 240×. Accessible flanking regions had an average length of 150 bp and comprised 0.55 Mb of accessible sequence. Regions flanking target sequence but interrupted by poor coverage at a small number of nucleotides, and more remote off-target regions, accounted for 0.29 Mb (13.5%) of the accessible genome.

DNA sequence variants

A total of 129,156 putative sequence variants (SNPs, MNPs and indels) were identified in the accessible genome (Table 6, XLS-file S3, and VCF-file S1). The density of substitution variants (SNPs and MNPs) was 1.6 times higher in non-coding regions than in coding regions, and the indel density was 12 times higher in non-coding regions. The transition/transversion ratio (T_s/T_v), calculated only for biallelic SNPs, was 1.55 and the ratio of non-synonymous to synonymous SNPs (pN/pS) was 0.64. Across the 84 cultivars, we observed 50 variants

generating premature stop codons in 43 genes. In addition, we found 130 indels expected to cause frameshifts in 90 target genes. A substantial fraction of these indels, however, was located near other indels that restored the coding frame via a second, compensatory frameshift.

TABLE 6. Overview of DNA variants observed across 84 cultivars in the accessible potato genome.

	Number of sequence variants called		
Variant type	*Accessible genome (2136 kb)*	*Non-coding (1480 kb)*	*Coding (656 kb)*
Di-nucleotide SNPs	105,812	84,454	25,358
Tri-SNPs	5,304	4,097	1,207
Tetra-SNPs	96	66	30
Indels	13,094	12,641	453
MNPs	4,850	4,084	766
Total	**129,156**	**101,342**	**27,814**

Across all cultivars, an average variant density of 1/24 bp in coding regions and 1/15 bp in non-coding regions was observed. Within a single tetraploid cultivar, on average 52,233 sequence variants (1/42.5 bp) were observed, generating an average of 116 cultivar-unique variants, ranging from 0 to 2,688. Cultivars like Vitelotte Noir, the only cultivar with purple flesh colour in our samples, and those with wild species introgression segments contained a relatively high number of cultivar-unique variants (e.g., up to 2.0% of all variants within cv. Vitelotte Noir). Cultivars without unique variants either had ancestors that were widely used in breeding of novel potato cultivars or had themselves been used for this purpose (e.g., cv. Agria and cv. Katahdin).

To evaluate the increase in variant density per additional sequenced cultivar, we permuted the order of the cultivars a thousand times and calculated the variant frequency at each incremental step (Figure 4). More than half of all variants were detected by selecting three random cultivars. When 16 random cultivars were selected, 84% of all variants were detected, and the number of novel variants that could be identified by sequencing an additional cultivar dropped below 1% of the variants already discovered. To detect 95% of all the identified variants, 46 random cultivars were required.

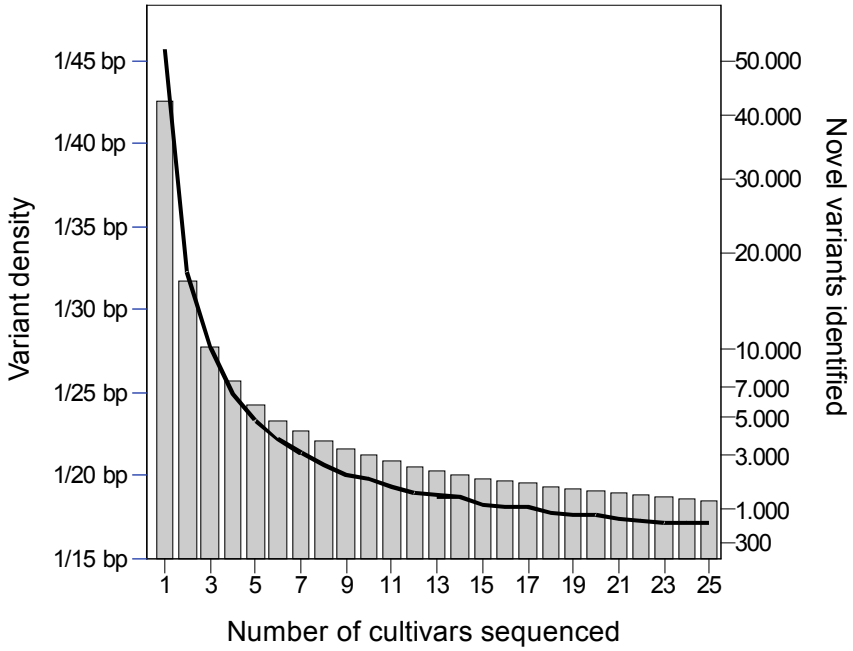

FIGURE 4. Sequence variant density as the number of randomly-added cultivars increases. The bars show variant density (primary Y-axis), and the black line shows the number of newly-identified variants (secondary Y-axis) as a function of the number of sequenced cultivars . Data is not shown after the 25th cultivar, but continues to drop to a variant density of 1/16.4 bp and an average of 116 novel variants at the 84th cultivar.

Genotyping and allele frequencies

To be assigned a valid genotype call for a specific, previously-identified variant position, a cultivar required a minimal read depth of 15× at that position. In total, 86.6% of all possible genotypes were valid (i.e., of the matrix of 84 cultivars by 129,156 sequence variants, 13.4% had insufficient read depth to make a call), equivalent to a per-locus average of 73 out of 84 cultivars receiving a genotype call. For 42,625 sequence variants (33% of all identified sequence variants), all 84 cultivars were genotyped, and more than 90% of all identified sequence variants were genotyped in at least half of the cultivars. Population-level allele frequencies for genotyped sequence variants were calculated using all valid calls. For 13,458 sequence variants, 10.4% of all identified, the allele in the DM reference genome differed from the major allele in the population. The distribution of minor allele frequencies (MAF) is shown in Figure 5. The average MAF was 0.14; 17.4% of all sequence variants had MAF<0.01, and 60.9% had MAF>0.05.

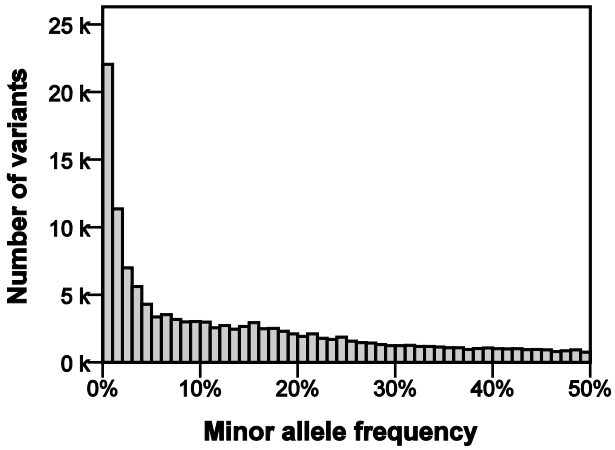

FIGURE 5. Distribution of minor allele frequencies (MAF) of all 129,156 genotyped sequence variants.

Nucleotide diversity

We next calculated the nucleotide diversity (π) for each of the 2,445 covered contigs, for each gene, and for each chromosome. Mean π of the covered genome was 1.07×10^{-4} (Table 5). As expected due to functional constraints on evolutionarily tolerable mutations, π was lower for coding regions than for non-coding regions (Table 5 and Figure 6, Panels A&B). Mean nucleotide diversity for chromosomes 5 and 11 was significantly higher overall, and individually at coding and the non-coding regions, relative to other chromosomes (Figure 6, Panels C-E). In contrast, the mean π overall for chromosome 10 was significantly lower than for other chromosomes, although mean nucleotide diversity was comparable to that of other chromosomes within coding regions. The physical position of contigs on the potato pseudomolecules was used to plot π over the twelve potato chromosomes (Figure 6F). Nucleotide diversity was low across most of chromosome 10, but it was comparable to other chromosomes near both ends. Individual genes with low (<0.006) nucleotide diversity were observed on all chromosomes (Table 7 & XLS-file S4). Individual loci lacking polymorphisms in both coding and non-coding regions were not detected, but several genes without any sequence variation in the covered coding regions were found (Table 7). These included important structural proteins, such as Histone H2B.11 and Axi1 protein.

TABLE 7. Examples of highly conserved potato genes. The two genes with the lowest nucleotide diversity per chromosome are shown. Genes in bold were 100% conserved in the coding region covered by resequencing. The nucleotide diversity of all covered genes is available in XLS-file S4.

		Covered length (bp)			Nucleotide diversity (π)			
Chr.	*Gene ID (PGSC0003)*	*Overall*	*Coding*	*Non-coding*	*Overall*	*Coding*	*Non-coding*	*Annotation*
1	DMG400000138	3,565	1,065	2,500	0.0025	0.0019	0.0027	Conserved gene of unknown function
1	DMG400026030	3,995	1,467	2,528	0.0029	0.0015	0.0036	Mitochondrial elongation factor
2	DMG400016714	921	466	455	0.0042	0.0031	0.0052	1-aminocyclopropane-1-carboxylate oxidase
2	DMG401010056	7,342	2,460	4,882	0.0044	0.0038	0.0047	CONSTANS 3 (StCO)
3	DMG400002066	442	263	179	0.0009	0.0010	0.0008	34 kDa outer mitochondrial membrane protein porin
3	**DMG400010377**	**279**	**103**	**176**	**0.0015**	**0.0000**	**0.0024**	**Axi 1 protein**
4	DMG400023995	3,256	2,533	723	0.0020	0.0018	0.0026	Trehalose-6-phosphate synthase
4	DMG400009936	1,299	701	598	0.0035	0.0009	0.0064	Beta-fructofuranosidase
5	DMG400033659	914	648	266	0.0043	0.0038	0.0056	SNF2 domain-containing protein / helicase domain-containing protein / RING finger domain-containing protein
5	DMG400023429	314	314	0	0.0056	0.0056	NA	Actin-58
6	**DMG400029167**	**445**	**376**	**69**	**0.0011**	**0.0000**	**0.0074**	**Histone H2B.11**
6	DMG400007122	923	852	71	0.0033	0.0034	0.0017	Arogenate dehydratase
7	DMG400018745	276	252	24	0.0027	0.0024	0.0049	Auxin efflux facilitator
7	DMG400017276	776	776	0	0.0046	0.0046	NA	Trehalose synthase
8	DMG400017702	1,342	554	788	0.0023	0.0002	0.0038	Methylenetetrahydrofolate dehydrogenase
8	DMG400003918	1,282	1,017	265	0.0029	0.0028	0.0031	Acc synthase
9	DMG400019345	1,129	1,129	0	0.0026	0.0026	NA	Ethylene overproducer-like 1
9	DMG400019316	884	150	734	0.0049	0.0060	0.0046	Homocysteine s-methyltransferase
10	**DMG402011945**	**333**	**171**	**162**	**0.0001**	**0.0000**	**0.0002**	**NTGP3**
10	DMG400024347	1,173	738	435	0.0010	0.0014	0.0004	Cytokinesis negative regulator RCP1
11	DMG402010918	284	194	90	0.0046	0.0010	0.0124	Xyloglucan endotransglycosylase
11	DMG400000439	269	269	0	0.0049	0.0049	NA	Actin-11
12	DMG400046738	410	317	93	0.0017	0.0022	0.0000	Gene of unknown function
12	DMG400004975	291	260	31	0.0052	0.0056	0.0019	Phosphopentothenoylcysteine decarboxylase
0 [a]	**DMG400007782**	**370**	**237**	**133**	**0.0008**	**0.0000**	**0.0021**	**Alpha-1,4 glucan phosphorylase L-1 isozyme, chloroplastic/amyloplastic**
0 [a]	DMG400006702	546	394	152	0.0024	0.0015	0.0047	Photosystem I reaction center subunit V, chloroplast

[a] unmapped superscafolds

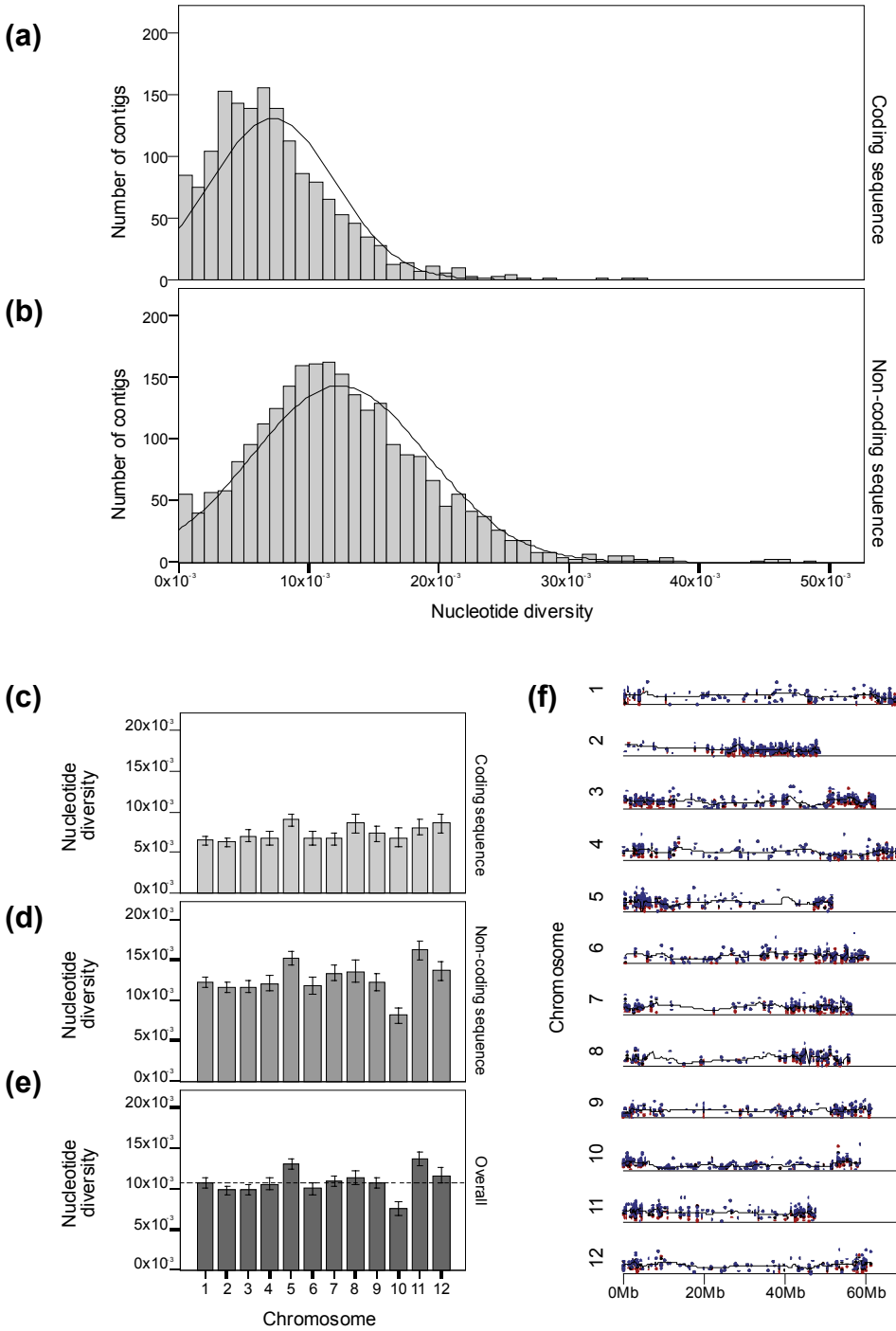
(a)
Number of contigs
Coding sequence
(b)
Non-coding sequence
0x10⁻³
10x10⁻³
20x10⁻³
30x10⁻³
40x10⁻³
50x10⁻³
Nucleotide diversity
(c)
Nucleotide diversity
Coding sequence
(d)
Non-coding sequence
(e)
Overall
Chromosome
(f)
Chromosome
0Mb
20Mb
40Mb
60Mb
80Mb

FIGURE 6. Nucleotide diversity of the accessible genome. Distribution of nucleotide diversity is shown for sequenced coding (A) and non-coding (B) regions. Bars represent the count of covered contigs and black line represents a normal distribution curve in panels A and B. Average nucleotide diversity per chromosome is shown for (C) coding, (D) non-coding, and (E) all sequenced regions. Dashed line represents overall genome mean nucleotide diversity (both coding and noncoding sequences), and error bars represent 95% confidence intervals in panels C, D, and E. (F) Nucleotide diversity ($\pi \times 10^{-3}$) of contig segments (average size 874±656 bp, mean±SD) scanning across chromosomes ordered in accordance with current pseudomolecule order. Nucleotide diversity at coding, non-coding, and all regions are coloured in red, blue, and black, respectively, while the fitted line represents the trend of nucleotide diversity fitted by 10% of contigs.

Population structure

Population structure was analysed using the 42,625 sequence variants that were genotyped in all 84 cultivars. The first three components of a principal component analysis described respectively 6.4%, 4.5% and 3.8% of the variance. In the centre of the PCA plot of the first two components, a cluster of cultivars of diverse origin were located (Figure 7). Three groups diverging from this set of cultivars were also noted, which consist of (a) heirloom British cultivars, (b) a number of typical frying cultivars from continental Europe, and (c) progenitors of potato cyst nematode (PCN) resistance and cultivars bred for starch industry where PCN resistance is necessary. The heirloom cultivar group is most similar to the more distant monoploid *S. tuberosum* Group Phureja clone, which fell into a cluster alone.

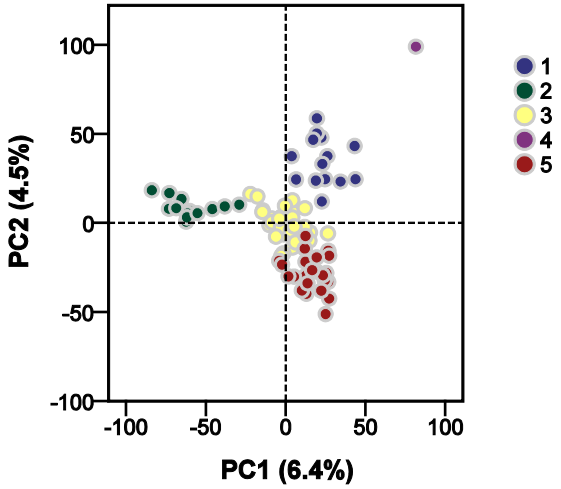

FIGURE 7. First and second components from principal component analysis of potato sequence variant genotypes. Population structure was analysed using ~43K sequence variants genotyped in all 84 cultivars. The first three components describe 14.7% of the variance. Based on these three components, the cultivars were clustered into five groups. The most distant cultivar is the monoploid *S. tuberosum* Group Phureja clone (Group 4)**.** In the centre of the PCA plot cultivars of diverse, world-wide origins are observed (Group 3). Three additional divergent groups can be observed, consisting of heirloom cultivars (Group 1), cultivars from continental Europe (Group 5) and cultivars and germplasm used in starch industry (Group 2).

Association analysis

The genotype dataset was also used for a genome wide association study (GWAS) to validate the sufficiency of this data for identifying known QTLs for plant maturity and tuber flesh colour. Association analysis was performed using both additive and dominant genotype models with separate tests with and without correction for population structure. Given the large number of low frequency alleles, the results of the dominant and additive models were largely consistent. The results of the dominant allele models, uncorrected for population structure, are shown in Figure 8.

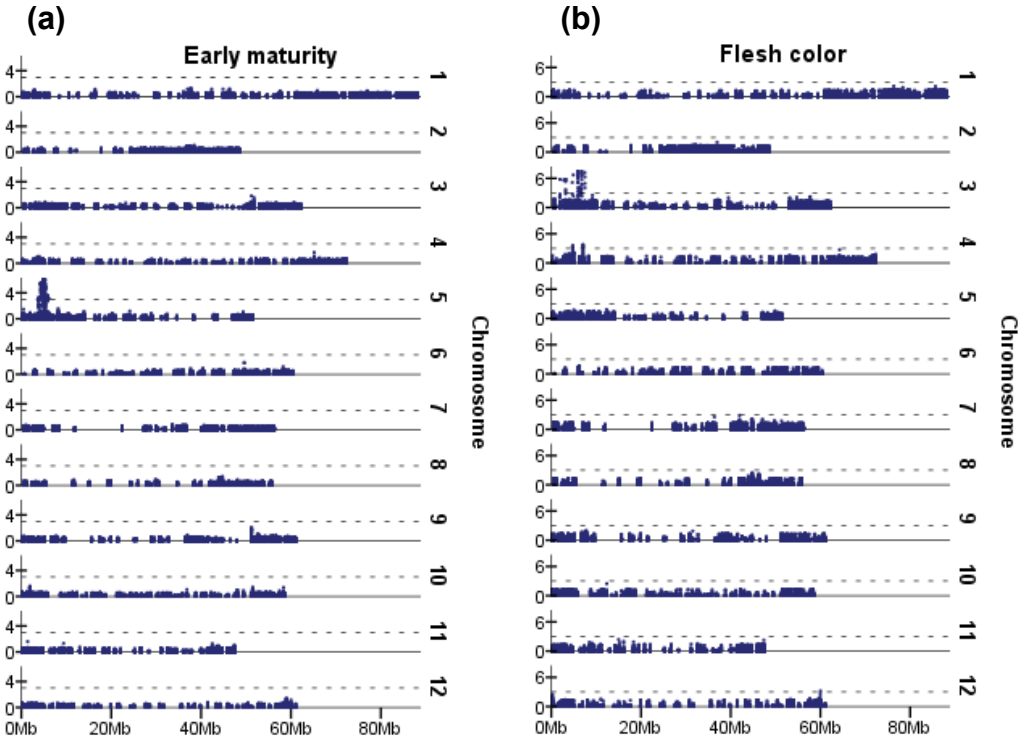

FIGURE 8. Manhattan plots of *p* values for associations between DNA sequence variants and two phenotypic traits in potato: (a) plant maturity and (b) tuber flesh colour. The FDR corrected -$\log_{10}(p)$ values from GWAS analysis are plotted against the physical position on each of 12 potato chromosomes. The horizontal dashed line is plotted at the FDR-corrected significance threshold of α=0.001 (-$\log_{10}(p)$=3).

Even with the small population size of 83 phenotyped cultivars, the well-known QTL for early plant maturity on potato chromosome 5 was clearly detected (-$\log_{10}(p)$=6.0; r^2 =0.44). Using the current data, this QTL mapped to a region of approximately 371 kb containing 28 strongly associated variants (-$\log_{10}(p)$≥5). For tuber flesh colour, a major QTL was observed on chromosome 3, mapping to a region of approximately 683 kb containing 27 strongly associated variants (-$\log_{10}(p)$ ≥7). These sequence variants were located within and near *CHY2* (*β-carotene hydroxylase*), a well-known gene influencing flesh colour via carotenoid synthesis, and

collectively explained 61% of the phenotypic variance. Two additional minor QTLs for flesh colour were found on chromosomes 4 and 12. In the model corrected for population structure, the QTL on chromosome 12 was not significant. The flesh colour QTL on chromosome 4 ($-\log_{10}(p)$=3.7) explained 9% of additional phenotypic variance beyond that explained by the major QTL. For both the plant maturity and the tuber flesh colour major QTLs, the relation between the allele-copy number of the best associate markers and the phenotypic effect could support either an additive or a dominant QTL model (examples provided in Figure 9).

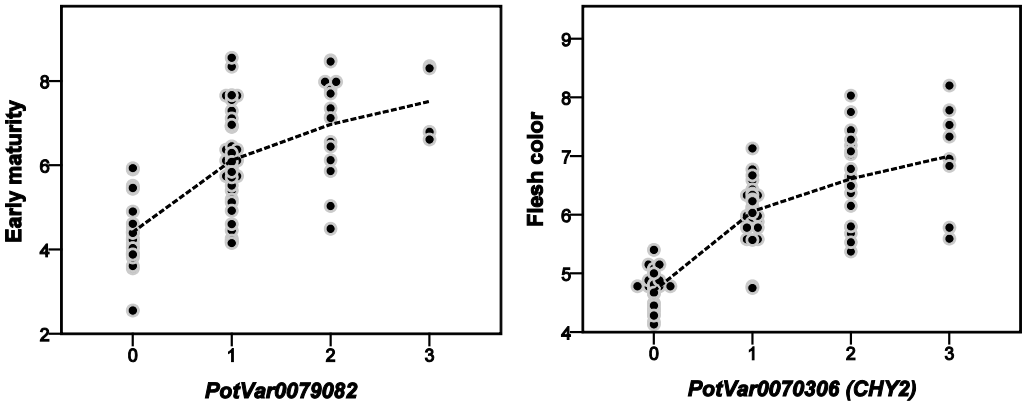

FIGURE 9. Scatterplots of early plant maturity and flesh colour phenotypic values against zygosity (allele copy number) of a strongly associated DNA sequence variant. For plant maturity, phenotype scores were enumerated from 1 (very late) to 9 (very early), and for flesh colour, scores ranged from 4 (white) to 9 (very yellow).

Validation of genotyping-by-sequencing data

A subset of 270 binary SNPs was selected to validate genotype calls made by GBS using KASP genotyping assays. When all GBS calls for this subset were scored as either homozygous (i.e., nulliplex or quadruplex) or heterozygous (i.e., simplex, duplex, or triplex), 97.9% of the homozygous and 99.9% of the heterozygous calls were concordant with the results of KASP genotyping. When heterozygous GBS calls were split into simplex, duplex and triplex categories, their overall concordance with KASP data dropped to 94.4%. For duplex calls, which were most difficult to classify, only 90.3% of the calls were concordant. This is not unexpected, since discrimination among simplex, duplex, and triplex genotypes requires read depths exceeding the 15× threshold. We therefore applied an additional genotype quality (GQ) filter to validate GBS/KASP concordance in the five discrete zygosity classes. GQ was calculated using a likelihood estimate to determine the probability that a genotype call was different from the true genotype. It was encoded as a phred quality score ($-10\times\log_{10}(p)$) and included read depth at a variant position as a parameter. A threshold of GQ≥26 was required to reduce the number of discordant duplex calls below 5%. Overall concordance between the GQ26 filtered set and KASP genotyping was 98.4%, with 96.2% concordant duplex calls (Table 8). Applying the GQ26 filter to the complete set of 129,156 sequence variants yielded 74.8% of

variants with an assigned genotype call, with an average of 63 out of 84 cultivars genotyped per variant position. This is equivalent to 25.2% genotypes not assigned, yielding approximately two-fold more unassigned genotypes compared to the set subject only to the 15× minimum threshold. Median read depth of all GBS genotype calls meeting the GQ26 threshold was 61×, and for duplex calls this was 81×.

TABLE 8. Concordance between genotyping-by-sequencing and KASP genotyping calls. The genotype calls derived from each method for 270 binary SNPs were compared. Sequencing calls were filtered by a minimum read depth of 15× and a minimum genotype quality score of GQ26.

	Expected genotype call (KASP)					
Genotyping-by-sequencing	*Nulliplex*	*Simplex*	*Duplex*	*Triplex*	*Quadruplex*	***Total***
Concordant calls	4,975	4,861	2,204	1,373	1,272	14,685
Discordant calls	67	25	86	34	25	237
Percentage of concordant calls	98.7%	99.5%	96.2%	97.6%	98.1%	98.4%
Percentage of alternative reads for concordant calls	0%	23%	49%	76%	100%	-
Percentage of alternative reads for discordant calls	4%	36%	34%	56%	97%	-

Analysis of chloroplast reads

As an initial analysis of chloroplast reads, 100,000 paired-end reads per cultivar, with index sequences on both sides, were mapped to the chloroplast reference genome. A total of 241 sequence variants, covering the complete chloroplast genome, were identified (Figure 10 and VCF-file S2). Since chloroplast sequences are monomorphic, haplotypes could be directly inferred; four main chloroplast types, with a number of sub-types, were identified here using a phylogenetic approach (Figure 11 and Table S1). A number of cultivars contained distinct chloroplast genomes resembling those of *S. demissum*, *S. vernei* and *S. tuberosum* Group Phureja, but most resembled those found in neo-tuberosum cultivars, such as cv. Desiree.

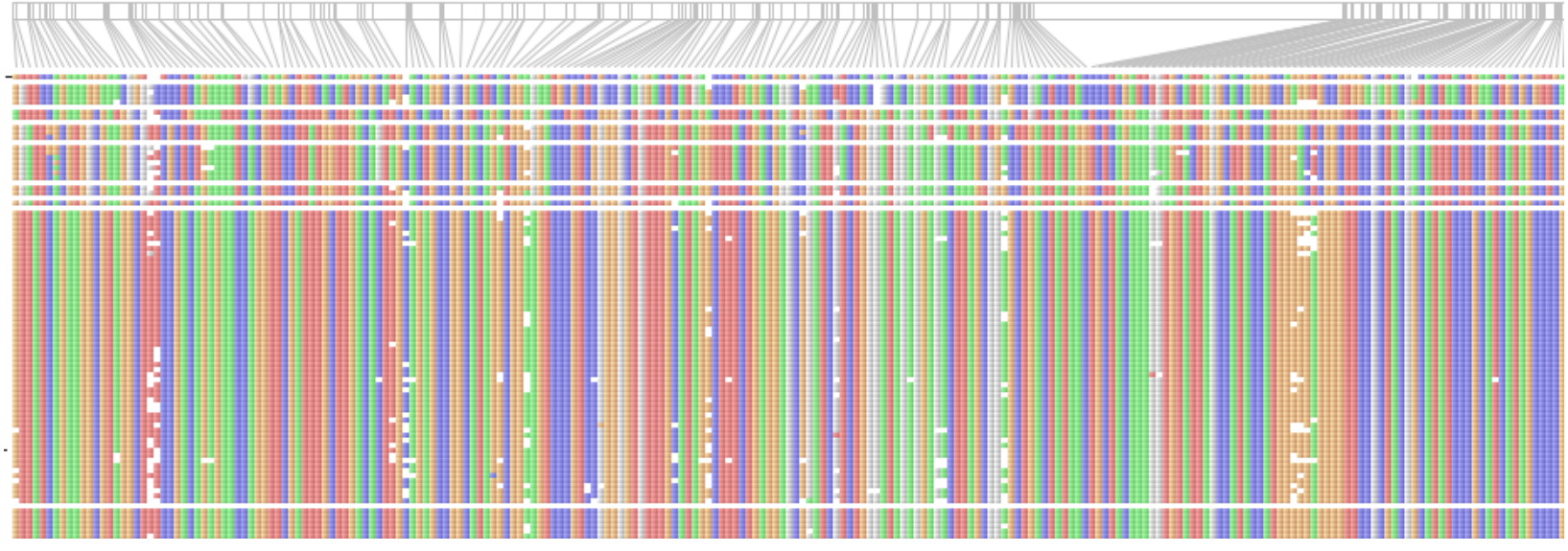

FIGURE 10. Alignment of chloroplast haplotypes. The top of the figure shows the complete potato chloroplast genome with the SNP positions. Cultivars are grouped according to similar haplotypes and for each cultivar the haplotype is given. Nucleotides at sequence variant positions are shown in green (A), red (T), blue (C) or yellow (G).

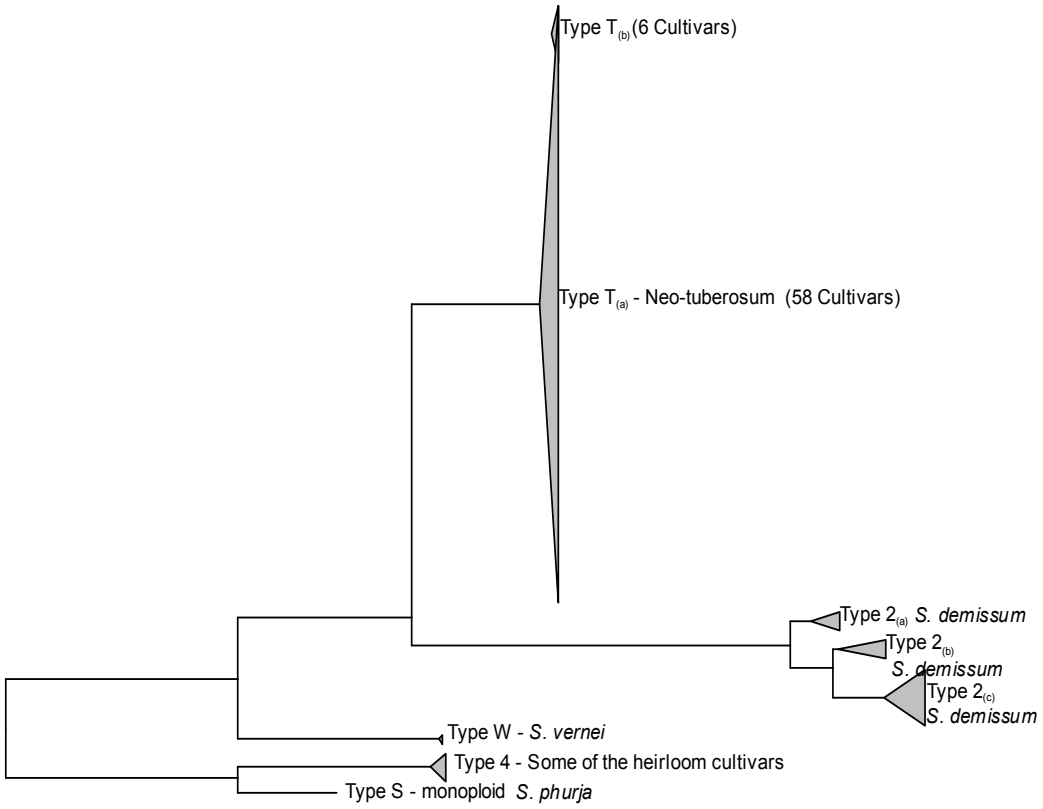

FIGURE 11. Neighbour-Joining tree of chloroplast haplotypes. The distances of 241 sequence variants for the 84 cultivars were computed using the Jukes-Cantor method and the tree inferred using the Neighbor-joining method.

DISCUSSION

Target-enriched genome sequencing

We used a mapping approach for aligning resequencing data with the recently sequenced potato DM genome (XU *et al.* 2011). A substantial minority of the reads (20%) could not be mapped to the reference genome. Since only initial chloroplast read filtering was completed here, most of the unmappable reads are presumed to be of chloroplast and mitochondrial origin. Assembly and/or alignment to chloroplast and mitochondrial reference genomes should provide more insight into the origin of the unmappable reads, but this is beyond the scope of the current study. The large share of cpDNA reads generated in this study (60% of total reads) was unforeseen in view of the 4% share of cpDNA baits used for target enrichment. In retrospect, this large proportion is probably due to copy number variation between nuclear DNA and cpDNA. The latter is present in about 100 copies per plastid and about 100 plastids per leaf cell (DANIELL 2004).

Of the sequences mapped to the reference genome, 45% aligned to target sequences. Other in-solution DNA enrichment studies in plant species have not yet been published, but this on target percentage is consistent with the 40-50% reached in human and animal studies that have used the SureSelect system for target enrichment (GNIRKE *et al.* 2009; HARISMENDY *et al.* 2009). We found consistent enrichment across all indexed samples, and virtually all target sequences were covered at the desired depth, with median average read depth of target sequences and the extended "accessible" genome of 88× and 63× per cultivar, respectively. Variation in read depth per sequenced cultivar was considerable, but this is attributable to repeated sequencing of a number of pools that had technical difficulties. Within each multiplexed pool, cultivar-specific read counts rarely reached a twofold difference between cultivars. The custom index adapters have proven valuable as multiplex adapters; no index-specific bias in read counts was observed and over 96% of the reads generated in this study were assigned to cultivars.

Sequence diversity analysis

This thesis chapter confirms the high sequence diversity of potato on a larger sample size and genome-wide scale than ever before. Compared to other crops where both coding and noncoding regions were analyzed for nucleotide diversity, total nucleotide diversity in potato (π=10.7×10^{-3}) is larger than that in sugar beet (7.6×10^{-3}, (MCGRATH *et al.* 2004), maize elite lines (6.3×10^{-3}, (CHING *et al.* 2002), and soybean (1.25×10^{-3}, (ZHU *et al.* 2003). We found an overall frequency of one variant every 17 bp in our population, 1/15 bp in non-coding and 1/24 bp in coding regions. This is somewhat higher than the one variant per 21-23 bp found in previous studies on potato (RICKERT *et al.* 2003; SIMKO *et al.* 2006). Comparison of nucleotide diversity between the current study and that by Simko et al. (2006) , with overall π=10.5x10^{-3} and π=14.6x10^{-3}, respectively, reveals somewhat lower nucleotide diversity in both coding and non-coding sequences for our study. Simko et al. (2006) sampled both fewer potato genotypes and fewer loci.

To investigate whether we could identify potential signatures of selection in the potato genome, we examined the nucleotide diversity of loci along the physical position of each chromosome. Nucleotide diversity values were extremely variable among individual, whole chromosomes. At the whole-chromosome level, chromosomes 5 and 11 exhibited highest nucleotide diversity. Introgression of resistance genes from wild species is a likely explanation, as both chromosome 5 and 11 contain large clusters of resistance genes conferring resistance to a wide variety of pathogens are present (BALLVORA *et al.* 2002; HUANG *et al.* 2005). In contrast, overall nucleotide diversity was reduced on chromosome 10. Reduced recombination and distorted segregation has been observed for chromosome 10 previously (BONIERBALE *et al.* 1988), and would be expected to result in reduced nucleotide diversity for the chromosome overall. This may also be related to selection for traits like general tuber impression and tuber shape, which has a major QTL on chromosome 10 (VANECK *et al.* 1994). Some of the most conserved genes in the dataset are located near the putative position of the tuber shape QTL on chromosome 10. Individual genes with a reduced nucleotide diversity are however clearly detected on all chromosomes. It is not possible at this stage to say whether

these genes themselves, rather than closely linked loci are under selection, but some of these genes are good candidate genes for phenotypic traits under strong selection like day-length dependent tuberisation and some resistance traits. An example of a good direct candidate gene is the *CONSTANS* gene that has been shown to affect the day-length regulation of tuber induction in potato (GONZÁLEZ-SCHAIN *et al.* 2012) and is the second most conserved gene we sampled on chromosome 2.

An important feature affecting the application of sequence variants as molecular markers is their minor allele frequency (MAF), which influences the type of information provided by the marker in different populations. Moderate-frequency alleles are valuable in applications where it is desirable to maximize the number of polymorphic markers between two parental lines. Low frequency alleles can, if not assayed, lead to missed lineages in phylogenetic reconstructions, overestimations of mean diversity (SCHLÖTTERER and HARR 2002), and spurious correlations in association mapping (PRITCHARD 2001). Here, we found an average MAF of 0.14, which is lower than for *Vitis vinifera* (average MAF of 0.24; LIJAVETZKY *et al.* 2007). In a population of 80 grapevine cultivars, over 80% of SNP variants had a MAF above 0.10, while in potato we found only 48% of the variants had a MAF above 0.10. It seems that a number of alleles, present in only one or a few cultivars, cause this high density of rare variants in potato. Most of these alleles are expected to be introgression segments. Owing to the largely vegetative mode of reproduction of potato, these segments are expected to be of considerable length and likely to cause linkage-drag.

For discovery of common sequencing variants, only a limited number of cultivars have to be sequenced, since more than half of all variants were already detected by sequencing three random cultivars. In a previous study in potato, SNPs have been identified in cDNA of a set of three to six – mainly North-American – potato cultivars (HAMILTON *et al.* 2011). In that study, SNPs were identified using Sanger EST-sequences available in GeneBank from three potato cultivars, and using high-throughput transcriptome sequencing on three additional cultivars (HAMILTON *et al.* 2011). With the relatively small sample size, these identified SNPs are expected to mainly represent the common coding SNPs present in the potato genepool. Compared to our study, a much larger part of the coding genome was however accessible for SNP calling. Overlap comparison of our data with the 69,586 mapped SNPs identified by Hamilton et al. (2011) indeed shows that only 2,572 of the latter are within our accessible sequenced genome regions. We detected 2,362 (92%) of these SNPs. As only one cultivar was included in both studies (cv. Bintje), the 8% of variants undetected in our study might represent rare variants more specific to the North American cultivars predominantly sampled by Hamilon et al. (2011), or may be false positive/negatives in either study. As a result of the at least 14-fold larger sample size and larger geographic diversity targeted by our study, we find approximately 15,000 extra variants in the exon sequences covered by both studies. These include additional multi-allelic sequence variants at 6% of the base positions where SNPs were called by Hamilon et al. (2011). When these SNPs are assumed bi-allelic, they may generate interference in genotype calling in for example a SNP genotyping array.

Population structure and chloroplast types

The population structure seen among the genotypes in this study largely coincides with that found for the same population using AFLP markers (D'HOOP *et al.* 2010). While the previous study distinguished five divergent cultivar groups, we identified four, with one outlier forming its own group. Both analyses support genetic similarity within heirloom cultivars and within starch cultivars, but the high-throughput approach used here did not support the presence of three distinct groups of fresh consumption, processing (frying), and additional miscellaneous cultivars. Instead, PCA analysis suggested the existence of a cluster of cultivars originating from continental Europe, consisting mainly of frying cultivars, along with a cluster containing cultivars of mixed, world-wide origin. Given that the first three components of the principal component analysis account for <15% of the total variation, it however seems that there is little the population structure within the potato cultivar genepool, as was also observered by D'hoop et al. (2010). With the heirloom cultivar group being most similar to the outlier clone in the PCA analysis, monoploid *S. tuberosum* Group Phureja, it seems that modern cultivars have diverged from the *S. tuberosum* Group Phureja material.

The presence of an extreme cytoplasmic bottleneck in cultivated potato has been known for a long time (PROVAN *et al.* 1999). Previous analysis has identified five main cpDNA types (A, S, C, W and T) and a number of sub-types (HOSAKA *et al.* 1988). The A haplotype is most frequent in Group Andigina and the T haplotype in Chilean S. *tuberosum* and modern cultivars. Diploid *S. tuberosum* Group Phureja is assigned to the S-type and wild material like *S. vernei* to the W-type. Our phylogeny of chloroplast haplotype data supports previous work suggesting that most modern cultivars have chloroplasts resembling those of the neo-tuberosum genepool (T-Type) (HOSAKA 2004; SPOONER *et al.* 2005). Two cultivars have a W-type haplotype originating from *S.vernei* and few cultivars have a chloroplast type originating from *S. demissum*. These chlorotypes have been introduced during introgression of resistance traits. More remarkably, four heirloom cultivars in our sample (cv. Belle De Fontenay, cv. Kepplestone Kidney, cv. Home Guard, and cv. Shamrock) have chloroplast haplotypes phylogenetically close to that from the *S. tuberosum* Group Phureja monoploid (S-Type). This chlorotype might represent the *S. tuberosum* Group Andigina type (A-Type) that was more common in cultivars from before the 1840s late blight (*Phytophtra infestance*) epidemic (POWELL *et al.* 1993).

Concordance of genotyping-by-sequencing and KASP genotyping in an autotetraploid species

Variant detection and genotyping is known to be highly dependent on sequence coverage and to be more difficult at heterozygous sites (KENNY *et al.* 2011; TEWHEY *et al.* 2009). Genotyping variants in polyploid species, such as potato, is even more challenging than in diploids, because a given gene may be represented not only by a number of different alleles, but also with different zygosity (nulliplex, simplex, duplex, triplex, and quadruplex). We therefore tested the accuracy of genotype calls made by GBS by genotyping a small subset of binary SNPs using an independent KASP genotyping platform. The genotype calls by GBS were found to be over 99% consistent with KASP when scored as either homozygous or heterozygous. For genotyping potato quantitatively in its five possible zygosity classes, a

higher read depth was required than our initial standard of 15× to achieve >95% consistency. We applied a genotype quality (GQ) parameter to account for differences in read depth requirement across zygosity classes. Less than 5% of duplex calls (the zygosity class with lowest consistency between GBS and KASP) were inconsistent between the two genotyping methods when GBS data were filtered by a genotype quality threshold set to GQ26, compared to ~10% using only the 15× depth requirement. Overall concordance rate of all five zygosity classes was 98.4% using the genotype quality threshold, while the number of non-scorable genotype values was doubled, to 25% of the total variants identified. Given the median read depth of the GQ26 filtered duplex calls in this study, in we recommend a median read depth of 80× for genotyping potato.

In the estimate of allele copy number we assume no ascertainment bias; i.e., our copy number data presumes the relative number of allele-specific sequencing reads is proportional to the zygosity. However, in-solution hybridisation enrichment may bias the pool of captured DNA targets that are ultimately sequenced towards those variants which preferentially hybridize and have higher sequence similarity to the reference bait sequence (GNIRKE *et al.* 2009). Anticipating this bias, we used a high probe tiling redundancy (6×, one RNA bait every ~20 bp) to reduce allele-specific bias during hybridization. Furthermore, we relaxed mapping quality settings for counting the reads mapped to the reference genome. In a preliminary genotype analysis performed after sequence variant calling, only reads with a very high mapping quality (≥MQ30) were used for determining the relative number of allele-specific sequencing reads. This caused a large ascertainment bias due to severe underrepresentation of non-reference alleles. Including reads with a lower mapping quality (MQ≥13) reduced this ascertainment bias, supporting the idea that hybridisation capture of more diverged alleles was efficient. We have not examined the genotyping accuracy of genetic variation from non-SNP sources, such as indels; the capture and mapping efficiency, and thus genotyping accuracy, may be lower for such sequence variants because they differ more from the reference sequence used for mapping and bait design.

Association analysis using GBS

The GBS dataset was robust in detecting common alleles influencing simple traits like plant maturity and tuber flesh colour using GWAS. The current samples size is however unlikely to reliably detect associations for more complex traits. Most associations we identified had similar statistical support in a dominant (present/absent) allele model as in an additive model. A dominant model may preferred in potato, given that around half of the identified sequence variants have a MAF below 0.10. A variant that follows Hardy-Weinberg equilibrium (HWE) with a MAF of 0.10 is not expected to be found as homozygous (quadruplex) and expected to be found in triplex, duplex and simplex in 0.4%, 5%, and 29% of the cultivars respectively. In practice, most variants will thus either be present in simplex or absent, and a dominant (present/absent) allele model will be sufficient. This has the advantage that the input GBS data does not require the more stringent genotype quality filtering necessary for zygosity estimation that leads to an increase in missing data for the additive model.

The strategy underlying GWAS is to genotype enough markers across the genome so that functional alleles will likely be in linkage disequilibrium (LD) –i.e., non-randomly associated at distinct loci – with at least one of the genotyped markers. Since we re-sequenced only a fraction (2.1 Mb, 0.25%) of the 840 Mb potato genome, sequence variants found associated with traits are more likely to be in LD with the trait loci rather than directly associated. For potato, based on AFLP markers, it is estimated that LD decays to background levels at a distance of around 4 cM (D'HOOP *et al.* 2010). For the data generated here, both short- and long-range LD still needs to be analyzed. Although short-range LD could have already been analyzed, long-range LD requires the robust ordering of superscaffolds of the DM reference genome into physical chromosomes, preferentially in combination with an aligned high-resolution genetic linkage map. These maps and pseudo-molecules are currently being developed and future effort will quantify the levels of LD decay on both genetic (cM) and physical scale (bp). Our expectation however is that the regions of the genome covered here are sufficient for most functional alleles to be in LD with a discovered sequence variant. Since a large number of genes in our study were either known candidate genes, or primary- and secondary metabolism genes, the annotated biochemical pathway may suggest a biological link between the gene underlying the sequence variants, and the trait. Alternatively, genes nearby associated sequence variants can be surveyed in the potato genome browser to find more likely candidate genes for the trait.

Development of high- and low-density SNP assays

While the current study demonstrates GBS as a feasible genotyping method for tetraploid potato, an alternative to GBS currently typically applied to genotype the large sample sizes required in methods such as association analysis is the use of high-throughput SNP assays (GANAL *et al.* 2011; LIJAVETZKY *et al.* 2007). The SNPs we identified in this study are an excellent resource for the development of such assays, because they were sourced from a relatively large and diverse population of potato cultivars. Thus we can estimate the frequency at which SNPs identified here are likely to appear in populations including those cultivars we sampled. Since the presence of many flanking SNPs near the one used for genotyping can cause hybridisation bias, and since potato exhibits high SNP frequency, our data also allows the potential for improved SNP assay design. In particular, we have identified target SNPs that exhibit a clean sequence context (i.e., no additional sequence variants cover the genotyping probe or target SNP) among the population of potato cultivars we sampled. From this data, we have developed both high- and low-density SNP assays that account for the identified flanking SNPs and can be used for a wide variety of applications that have previously been discussed in detail for other crop species (MCCOUCH *et al.* 2010) and include marker-assisted and genomic selection, association and QTL mapping, positional cloning, diversity and pedigree analysis, and variety identification.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Björn D'hoop and Dr. Jan de Boer for help in the selection of AFLP target sequences and access to the phenotypic data. We thank Dr. Joao Paulo and Ir. Peter Vos for assistance in the marker-trait analysis. For bioinformatics support we thank Dr. Theo Borm. We are very thankful to Dr. Alexander Hoischen for allowing us access to the Covaris DNA shearing apparatus at Radboud University, and to Pieter van der Vlies of University Medical Centre Groningen for sequencing our libraries. This research was supported by a grant from the Dutch technology foundation STW, project WPB-7926.

SUPPLEMENTARY DATA

TABLE S1. Potato clones used in this experiment.

Name[a]	*Year of release*	*Origin*[b]	*Genotype Code*[c]	*Pool ID*	*Adapter Index ID*	*Mean readdepth*[d]	*Bases with depth>15x (%)*	*Population cluster*	*Chloroplast type*
1256A(23) = Black 1256	?	GB	P80001	P1	PEM01	25.60	65.0	1	Type T(a)
Arrow	2004	HOL	P80031	P1	PEM02	30.28	71.7	3	Type T(a)
Innovator	1999	HOL	P80106	P1	PEM03	56.94	89.5	3	Type T(a)
Vitelotte Noir	<1815	FRA	P80232	P1	PEM04	67.37	91.1	3	Type T(b)
Ve 71-105	1971	HOL	P80206	P1	PEM05	73.28	92.6	2	Type W
Aurora	1972	HOL	P80035	P1	PEM06	15.41	48.4	3	Type T(b)
Yam	<1787	GB	P80221	P1	PEM07	43.79	80.3	1	Type T(b)
Ultimus	1935	HOL	P80199	P1	PEM08	84.92	94.9	5	Type T(b)
Shamrock	<1900	IRL	P80182	P1	PEM09	41.75	79.8	1	Type4(a)
Princess	1998	BRD	P80234	P1	PEM10	65.25	89.5	5	Type 2(a)
Aveka	2001	HOL	P8B119	P1	PEM11	35.36	72.9	2	Type T(a)
Eos	2000	HOL	P80076	P1	PEM12	36.50	77.3	3	Type T(a)
Nomade	1995	HOL	P80151	P2	PEM01	80.72	90.3	2	Type T(a)
Markies	1997	HOL	P80139	P2	PEM02	78.43	90.0	5	Type T(a)
Adretta	1975	DDR	P80006	P2	PEM03	46.08	78.2	5	Type 2(c)
Kerpondy	1949	FRA	P80118	P2	PEM04	30.98	68.7	5	Type T(a)
Victoria	1997	HOL	P80208	P2	PEM05	62.26	81.9	5	Type T(a)
Biogold	2004	HOL	P80045	P2	PEM06	34.57	74.0	3	Type T(a)
Pentland Dell	1961	GB	P80159	P2	PEM07	14.65	46.6	1	Type T(a)
Toyoshiro	1976	JAP	P80192	P2	PEM08	43.22	73.1	3	Type T(b)
Kepplestone Kidney	<1900	GB	P80117	P2	PEM09	50.52	81.1	1	Type4(b)
Fontane	1999	HOL	P80088	P2	PEM10	67.81	83.7	5	Type T(a)
Felsina	1992	HOL	P80082	P2	PEM11	74.91	84.7	5	Type T(a)
Fianna	1987	HOL	P80084	P2	PEM12	64.53	85.8	3	Type T(a)
Industrie	1900	GER	P80105	P3	PEM01	62.41	90.8	5	Type T(a)
Samba	1989	FRA	P80176	P3	PEM02	60.89	87.4	3	Type T(a)
Libertas	1946	HOL	P80131	P3	PEM03	96.41	98.2	3	Type T(a)
Vk 69-491	1969	HOL	P80211	P3	PEM04	92.02	96.5	2	Type T(a)
Mpi 19268	1940	GER	P80145	P3	PEM05	80.15	93.9	5	Type 2(a)
Agria	1985	BRD	P80008	P3	PEM06	97.72	97.2	5	Type T(a)
Ajiba	1992	HOL	P80009	P3	PEM07	107.59	97.7	3	Type T(a)
Y 66-13-636	1966	HOL	P80220	P3	PEM08	127.26	97.8	3	Type T(a)
Herald	1928	GB	P80099	P3	PEM09	49.47	84.8	3	Type T(a)
Anosta	1975	HOL	P80022	P3	PEM10	145.05	99.1	3	Type T(a)
Ve 74-45	1974	HOL	P80207	P3	PEM11	99.76	96.2	2	Type T(a)
Tasso	1963	BRD	P80188	P3	PEM12	84.95	95.5	3	Type 2(c)
Nicola	1973	BRD	P80147	P4	PEM01	116.71	98.0	5	Type T(a)
Kuras	1996	HOL	P80122	P4	PEM02	163.07	99.3	2	Type 2(b)
Ehud	1965	HOL	P80073	P4	PEM03	98.06	98.1	3	Type 2(c)
Hindenburg	1916	GER	P80101	P4	PEM04	141.58	99.4	5	Type 2(ba)
Golden Wonder	1906	GB	P80095	P4	PEM05	104.55	98.2	1	Type T(a)

Name[a]	*Year of release*	*Origin*[b]	*Genotype Code*[c]	*Pool ID*	*Adapter Index ID*	*Mean readdepth*[d]	*Bases with depth>15x*	*Population cluster*	*Chloroplast type*
Vtn 62-33-3	1962	HOL	P80214	P4	PEM06	105.78	96.7	2	Type W
Avenance	2005	HOL	P80037	P4	PEM07	94.48	97.1	2	Type 2(c)
Ve 70-9	1970	HOL	P80205	P4	PEM08	137.6	99.1	2	Type T(a)
Hansa	1957	BRD	P80098	P4	PEM09	106.01	97.1	5	Type T(a)
Ackersegen	1929	GER	P80003	P4	PEM10	174.54	99.3	5	Type T(a)
Umatilla Russet	1998	USA	P80200	P4	PEM11	177.2	99.1	3	Type T(a)
Laura	1998	BRD	P80128	P4	PEM12	100.42	96.7	5	Type T(a)
Picasso	1994	HOL	P80161	P5	PEM01	73.59	90.5	3	Type T(a)
Obelix	1988	HOL	P80153	P5	PEM02	61.93	89.4	3	Type T(a)
Monoploid 1-3 551	NA	NA	NA	P5	PEM03	51.72	81.8	4	Type S
Mondial	1987	HOL	P80143	P5	PEM04	46.45	82.3	3	Type T(a)
Exquisa	1992	BRD	P80080	P5	PEM05	47.76	81.4	5	Type T(a)
Frieslander	1990	HOL	P80090	P5	PEM06	81.26	93.0	3	Type T(a)
Ballydoon	1931	GB	P80038	P5	PEM07	29.91	70.0	1	Type T(a)
Bintje	1910	HOL	P80044	P5	PEM08	90.97	94.3	3	Type T(a)
Wisent	2005	HOL	P80219	P5	PEM09	60.17	88.9	2	Type 2(b)
Irish Queen	<1900	GB	P80109	P5	PEM10	69.21	91.4	1	Type T(a)
Festien	2000	HOL	P80083	P5	PEM11	47.83	80.5	2	Type 2(c)
Hermes	1973	AUT	P80100	P5	PEM12	90.71	93.6	5	Type T(a)
Katahdin	1932	USA	P80115	P6	PEM01	73.63	93.5	3	Type T(a)
Clivia	1962	BRD	P80053	P6	PEM02	52.1	84.7	5	Type T(a)
Gladstone	1932	GB	P80093	P6	PEM03	33.7	72.6	1	Type T(a)
Ditta	1989	AUT	P80062	P6	PEM04	92.33	96.3	5	Type T(a)
Estima	1973	HOL	P80079	P6	PEM05	100.85	96.3	3	Type T(a)
Arran Pilot	1930	GB	P80029	P6	PEM06	32.27	72.0	3	Type T(a)
Tinwald's Perfection	1914	GB	P80191	P6	PEM07	65.09	90.2	1	Type T(a)
Albion	1895	HOL	P80010	P6	PEM08	50.40	85.3	3	Type T(a)
Amyla	1999	FRA	P80021	P6	PEM09	39.93	77.3	3	Type 2(a)
Alpha	1925	HOL	P80015	P6	PEM10	107.47	97.8	3	Type T(a)
Winston	1992	GB	P80218	P6	PEM11	45.5	79.4	3	Type T(a)
Great Scot	1909	GB	P80097	P6	PEM12	57.07	89.3	1	Type T(a)
Civa	1960	HOL	P80052	P7	PEM01	24.71	65.7	5	Type T(a)
Arran Chief	1911	GB	P80028	P7	PEM02	32.76	76.1	1	Type T(a)
Usda 96-56	?	USA	P80203	P7	PEM03	47.5	82	3	Type T(a)
Belle de Fontenay	1885	FRA	P80040	P7	PEM04	29.5	69.9	1	Type4(a)
Kartel	1994	HOL	P80114	P7	PEM05	45.37	83.5	2	Type 2(c)
Home Guard	1943	GB	P80102	P7	PEM06	36.29	73.4	1	Type4(a)
Daisy	1998	FRA	P80057	P7	PEM07	23.14	64.1	3	Type T(b)
Voran	1931	GER	P80212	P7	PEM08	60.06	87	3	Type T(a)
Cherie	1997	FRA	P80050	P7	PEM09	41.95	82.6	3	Type T(a)
Mercator	1999	HOL	P80141	P7	PEM10	77.38	93.4	2	Type 2(c)
Vivaldi	1998	HOL	P80210	P7	PEM11	69.19	91.2	5	Type T(a)
Charlotte	1981	FRA	P80049	P7	PEM12	32.81	75.5	5	Type T(a)

[a] Name of cultivar, progenitor clone, or monoploid clone
[b] Country of first market release
[c] Same as P8 codes defined by D'hoop et al. (2008)
[d] MQ13 of covered regions

SUPPLEMENTARY ELECTRONIC FILES

FASTA-FILE S1. **SureSelect RNA bait sequences.**

FASTA-FILE S2. **SureSelect target sequences.**

XLS-FILE S1. **Annotations for genomic SureSelect targets, including observed coverage in resequencing data.**

XLS-FILE S2. **Annotations for accessible genome regions, based on sequence data collected in this study.**

XLS-FILE S3. **Annotations for sequence variants identified in the accessible genome, including allele copy numbers of each of the 84 samples.**

XLS-FILE S4. **Annotations of nucleotide diversity for accessible genes.**

BED-FILE S1. **SureSelect baits mapped (MQ≥37) to the superscaffolds of the DM reference genome.**

BED-FILE S2. **Accessible genome regions of the DM reference genome.**

VCF-FILE S1. **Sequence variants and genotypes identified in the accessible potato DM genome of 84 samples.**

VCF-FILE S2. **Sequence variants identified in the chloroplast genome.**

CHAPTER 5

EMS-Induced Mutation Discovery in the M_1 Generation of Potato as a Strategy for Reverse Genetics

AUTHORS
J.G.A.M.L. Uitdewilligen
A.M.A. Wolters
H.J. van Eck
R.G.F. Visser

ABSTRACT

We applied Ethyl methanesulphonate (EMS) to diploid potato by two different mutagenic treatments and screened the resulting populations for novel mutations using high-resolution melting (HRM) analysis. A pollen-treatment with EMS dissolved in a sucrose solution induced mutations at a low frequency. *In planta* selection of the most vital mutagenized pollen seems to have lowered the mutation density to a frequency that is not suitable for reverse genetics studies. The EMS seed-treatment on the other hand provided a high density of novel mutations. In contrast to most EMS mutagenesis studies, we directly screened the M_1 generation of the seed-treated population. The high mutation density of 1/65 kb we found in the seed-treated population makes screening of the M_1 generation an attractive system for obtaining mutations. A large spectrum of 65 novel alleles for six candidate genes involved in starch metabolism was identified in the M_1 population. For all six genes, missense mutations that are predicted to damage protein function were discovered and for four genes premature stop codon mutations were identified. Genetically stable M_2 and M_3 plants have been generated for 10 of the most interesting mutations (37% of the original mutations). The estimated mutation density of M_1 mutations that are transferable to the M_2 generation (one "accessible" mutation/118-176 kb) is higher than the mutation density obtained by M_2 screening studies of most other plant species (KUROWSKA *et al.* 2011). The results thus demonstrate that M_1 screening offers a practical alternative to the commonly applied M_2 screening for the rapid generation of novel genetic variation at a high density, without too many complications in recovering mutations in the M_2 generation.

INTRODUCTION

Commercial potato plants are autotetraploid cultivars resistant to inbreeding and with high levels of heterozygosity and nucleotide diversity (see previous Chapters). Alleles identified by analyzing the natural variation present in a genepool of tetraploid potatoes have been filtered by (natural) selection. Although mutant alleles accumulate in polyploid populations more quickly than in diploids (OTTO 2007) and potato has a high genetic load (VAN ECK *et al.* 1994), spontaneous knockout or reduction-of-function mutations are expected to be relatively rare. Chemical mutagenesis with agents such as ethyl methanesulphonate (EMS) is a rapid cost-effective method for generating this kind of new genetic variation and can be used to unravel biological processes and for the alteration of agronomic traits. EMS treatment predominantly induces C-to-T and G-to-A DNA transitions randomly throughout the genome (SEGA 1984). It results in high point mutation densities with only low levels of chromosome breaks that for example cause aneuploidy, reduced fertility, and dominant lethality in atomic bomb and X-Ray irradiation (MOH 1950). In contrast to insertional mutagenesis like T-DNA or transposon/retrotransposon tagging that generate mostly knockouts, chemical mutagens like EMS can induce a series of alleles for a targeted locus. In addition to loss-of-function alleles, it generates alleles with reduced, enhanced or even novel gene function and thus can provide a range of alternative phenotypes (ALONSO and ECKER 2006). Furthermore, as chemical

mutagenesis requires no genetic transformation, it is widely applicable to both model species and crop species. EMS mutagenesis therefore has become the method of choice for many crop improvement and gene function studies and its application in reverse genetics has already been demonstrated for a large number of economically important crops, including rice, barley, wheat, maize, sorghum, soybean, rapeseed, tomato and potato (KUROWSKA *et al.* 2011).

The method of application of EMS and the genetic structure of the target populations can vary. Typically, EMS mutagenesis is carried out by soaking seed in an EMS solution. Using this method, both parental genomes are targeted. Individuals arising from the mutagenized seeds (the M_1 generation) are however chimeric. To avoid the chimeric tissue, commonly the M_2 generation of selfed M_1 plants is used for mutation screening. The requirement of this second non-chimeric M_2 generation for screening makes it time consuming. Furthermore, since induced mutations segregate in a M_2 population, of each M_1 plant multiple M_2 family plants, or seeds, have to be screened in order not to miss the segregating mutation. An alternative mutagenesis method extensively applied in maize is the EMS treatment of pollen. This method is described by Neuffer and Coe (1978) and is based on mixing paraffin-oil diluted EMS with pollen. It has been applied in Eucalyptus (MCMANUS *et al.* 2006) and been coupled to reverse genetic screening in maize (TILL *et al.* 2004). In EMS pollen-treatment, M_1 individuals produced by fertilization with mutagenized pollen have only one mutagenized genome and are non-chimeric. Therefore pollen-treated populations are screened at the M_1 generation (TILL *et al.* 2004). To determine which EMS doses permit efficient mutant generation, protocols describing mutagenic treatment of seed commonly recommend assaying seed mortality or seedling growth in response to treatment (IAEA 1977; MULLARKEY and JONES 2000). When pollen rather than seed is treated, assaying pollen germination reveals which doses do not completely render pollen non-viable, whilst changing pollen behaviour *in vitro*. This change is important because it indicates that pollen, and therefore M_1 seedlings, should carry induced mutations (MCMANUS *et al.* 2007).

The efficiency of a mutagenesis approach is defined as "the number of events per genome that are inherited in a population that has been mutagenized under standard conditions" (ALONSO and ECKER 2006). At equal mutation densities per mutagenized genome, only half the number of EMS seed-treated M_1 plants have to be screened in order to find a mutation compared to pollen-treated M_1 plants, since in seed-treated EMS populations both parental genomes are mutagenized and in pollen-treated populations only one. This makes seed-mutagenesis more efficient then pollen-mutagenesis. The number of mutagenized genomes per plant is usually however not considered when calculating mutation densities. Usually, to calculate a mutation density the total number of plants screened is multiplied by the length of the screened DNA target and divided by the number of identified mutations. Calculated in this way, a mutation density of 1 mutation/25 kb in seed-treated hexaploid wheat seems much higher than a mutation density of 1 mutation/40 kb in tetraploid wheat (SLADE *et al.* 2005). Mutation densities per mutagenized genome in this example are however very comparable; 1 mutation per respectively 150 kb and 160 kb of mutagenized genome. As an alternative, mutations rates can be calculated as the number of mutations found per 1 kb length screened in 1,000 plants. Also here, the number of mutagenized genomes is not considered but this measurements

gives a clear estimation on how many plants have to be screened to find a number of mutations.

The most established method for the detection of DNA polymorphisms in EMS treated populations is a heteroduplex mismatch cleavage assay based on the endonuclease CEL1 (McCallum *et al.* 2000). When using restriction-based mutation screening methods, a low level of natural polymorphisms in the gene of interest is however a requirement for the efficient detection of novel mutations. An alternative technology, High resolution melting (HRM) analysis, derived from the combination of existing techniques of DNA melting analysis with a new generation of fluorescent dyes (Wittwer *et al.* 2003) could also be used. HRM analysis is applied to analyse genetic variations including SNPs, length polymorphisms and methylation of DNA in PCR amplicons and has been applied successfully in a number of mutagenesis screens (Botticella *et al.* 2011; Bush and Krysan 2010; Gady *et al.* 2009; Ishikawa *et al.* 2010; Reed and Wittwer 2004). The time and costs of HRM are similar to conventional PCR while it avoids the need for post-PCR separation required by many other assays, making it a very efficient method for reverse genetics screens.

The first mutagenesis experiments in potato induced mutations by Röntgen-irradiation, and identified a loss-of-function mutation in the *GBSS* candidate gene by screening monoploid mini-tubers for absence of amylose starch (Hovenkamp-Hermelink *et al.* 1987). Reverse genetic mutation screening, where mutations are first identified at the DNA level, so far has only been conducted using an EMS treated dihaploid potato population, screened by direct Sanger sequencing for mutations in the same *GBSS* gene (Muth *et al.* 2008). In reverse genetics screening, a M_1 generation is usually selfed to obtain a M_2 generation that harbours homozygous mutants with potential phenotypes. In diploid potato gametophitic incompatibility systems are however active that complicate selfing (Eijlander *et al.* 1997). To obtain a genetically stable M_2 generation, Muth et al. (2008) used callus regeneration and spontaneous chromosome doubling to obtain a tetraploid M_1 clone with the mutation of interest, and crossed this clone with pollen of tetraploid donor plants. The M_2 plants were crossed amongst each other to obtain a homozygous mutant expressing the waxy phenotype of interest. Self-compatible diploid potato clones exist (Hosaka and Hanneman 1998; Olsder and Hermsen 1976) and might be useful as alternative to quickly obtain diploid M_2 plants homozygous for the mutation of interest for phenotypic evaluation. These mutants could be incorporated into diploid potato breeding programmes and introduced into tetraploid cultivars by 2n-pollen formation (Hutten *et al.* 1994).

In this chapter we test the efficiency of EMS-mutagenesis by pollen- and seed treatment in diploid potato and the density of genetically stable mutations, accessible in the next generation.

MATERIALS AND METHODS

Plant material and EMS mutagenesis

Pollen mutagenesis

Pollen of the highly self-compatible and fertile G254 potato clone – a dihaploid of *S.tuberosum* cv. Gineke (OLSDER and HERMSEN 1976) – and of RH89-039-16 (RH) (VAN OS *et al.* 2006) were tested for EMS pollen mutagenesis. G254 was chosen since this clone shows low photoperiod-sensitivity (long flowering time), is self-compatible and has a high pollen fertility (OLSDER and HERMSEN 1976). Disadvantages of this clone are its low seed fertility due to three lethal mutations (HERMSEN 1978) and its low agronomical value. Batches of approximately 110 mg pollen were mixed with 1.5 ml Ethyl methanesulfonate (EMS) solution with varying concentrations, buffered by 15% sucrose. Samples were incubated on a shaker/roller for one hour, pipetted in filter tubes (Ultra free-MC 0.65 μm, Millipore) and rinsed three times with a 15% sucrose solution (500 μl) by centrifuging on a table-top centrifuge for 4 minutes at the lowest speed (500 rpm, 27× g). After the last rinse the pollen was suspended in 80 μl 15% sucrose solution and applied for fertilisation of emasculated G254 flowers or for *in vitro* germination to estimate optimal EMS concentrations. For *in vitro* germination, 4 μl of the pollen solution was pipetted into 400 μl of germination solution (15% sucrose solution with 1.6 mM H_3BO_3) in a 24-well tissue culture plate, with two wells acting as a replicate for each treatment. After 20 h, samples were scored under a microscope for pollen germination percentage using a random sample of 200 pollen grains per replicate.

Seed mutagenesis

For determining the optimal EMS treatment concentration of seeds, batches of 150 mg of selfed G254 seeds (G254⊗, I_1) were used (approximately 320 seeds). To break dormancy and initiate germination, seeds were treated with a gibberellin solution (1.84 mg/mL) for 24 h on filter paper, washed and left for another 24 h on wet filter paper. The germinating seeds were added to 5 ml of EMS solution with varying concentrations of EMS diluted in water and incubated on a shaker/roller for 16 h. To inactivate the EMS, seeds were rinsed in a 3% Na-thiosulfate solution for 20 minutes and rinsed again 2 × 20 minutes in water. Seeds were sown in sowing trays and evaluated for seedling emergence after 20, 33 and 45 days after sowing. For generation of the bulk G254⊗ 0.6% EMS [v/v] population the treatment protocol was scaled up five times, using six batches of 750 mg of G254⊗ seeds (around 1600 seeds/batch).

Mutation detection and validation

Genomic DNA was isolated from leaf material using Kingfisher Genomic DNA Purification Kit (Thermo Hybaid) and the Kingfisher magnetic nucleic acid extraction system (Thermo Labsystems) according to the manufacturer's procedure. DNA concentrations were quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific). Amplicons for HRM genotyping were generated from 15 ng genomic DNA template per plant. PCR amplifications were performed in 10 μl reactions using 2 μl of F-524 Phire™ 5× reaction buffer (Finnzymes), 0.1 μl Phire™ Hot Start DNA Polymerase (Finnzymes), 1 μl LCGreen™ Plus+ (BioChem) and

0.25 µl of 5 mM primers (Table S1). PCR and heteroduplex formation were performed using the following conditions: 94°C, 2 minutes; 40 cycles, 94°C, 5 seconds; fragment-dependent Tm, 10 seconds; 72°C, 10 seconds; a denaturation step of 30 seconds at 94°C and renaturation by cooling to 30°C. Amplicons of either individual plants, or pools of four plants, were screened for mutations using the LightScanner® (Idaho Technology) HRM analysis system and the primer sequences described in Table. Amplicons of individual plants of putative mutant pools were re-screened using the LightScanner® HRM system. Plants with a putative mutation were sequenced on a ABI3700 sequencer using the dideoxy chain-termination method and ABI PRISM Reaction Kit. As sequencing primer, the forward amplification primer was used. To study possible effects of amino acid changes on functionality of the protein the programs SIFT (Sorting Intolerant From Tolerant) (NG and HENIKOFF 2003) and PolyPhen-2 (Polymorphism Phenotyping) (ADZHUBEI *et al.* 2010) were used.

In vitro multiplication and stabilisation of mutants

Internodal stem segments of chimeric M_1 plants carrying an identified mutation were surface-sterilized and *in vitro* cultured. Genomic DNA of these clones was isolated using the Kingfisher system and screened for the identified mutation using the Sanger sequencing and/or LightScanner® HRM system as described above. Clones carrying the mutation were used as maternal parent in crosses with RH89-039-16 (RH) (VAN OS *et al.* 2006). Presence of the mutation in the M_2 (M_1×RH) offspring was verified using Sanger sequencing and/or the LightScanner® HRM system. M_2 plants heterozygous for the specific mutation were mutually crossed to obtain the M_3 (M_2×M_2) generation and checked for the presence of the mutation using Sanger sequencing and/or the LightScanner® HRM system.

RESULTS

Development of EMS-induced mutant populations

Several EMS concentrations, incubation times and buffer solutions were tested to find the suitable parameters to treat G254 and RH pollen with EMS (data not shown). Best results were obtained by treating G254 pollen in a 15% sucrose-buffered EMS solution for one hour. *In vitro* germination showed a linear correlation between G254 pollen germination and EMS concentration in the 0-2% EMS range. Pollen germination was approximately half of the control (0% EMS) at a concentration of 1% EMS (Figure 1).

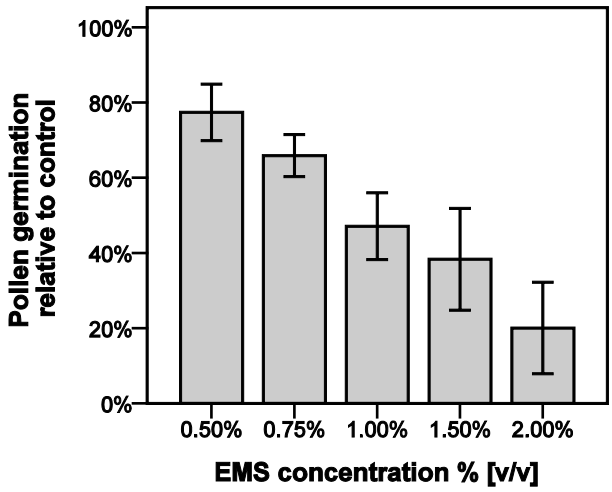

Figure 1. In vitro pollen germination of pollen treated with several EMS concentrations. Error bars represent ±2×SE.

To establish the pollen-treated EMS plant populations (total of 1,824 M_1 plants) we treated G254 pollen with three different EMS concentrations and used this pollen for selfing G254 plants (Table 1). The relative pollination success (number of berries per attempted cross) dropped below 50% of control at an EMS concentration of 1.5%. The number of seeds per berry dropped below 50% of control at a concentration of 1% EMS, while seed germination remained relatively high at all three concentrations. At the highest EMS concentration used for *in vivo* pollination, the M_1 seed germination was still significantly higher than the seed germination in the seed-treated EMS population (68% and 45% in respectively the pollen-treated and the seed-treated population).

TABLE 1. Pollen-treated EMS populations.

Pollen EMS concentration	*Attempted pollinations*	*Successful pollinations*[a]	*Seeds per berry*[a]	*Seed germination*[a]	*Plants used for DNA extraction and HRM screening*	*Tuber-forming plants*
0.63%	30	28 (100%)	76 (66%)	(85%)	414	75%
1.00%	147	92 (67%)	48 (42%)	(74%)	1269	67%
1.50%	69	31 (48%)	20 (17%)	(68%)	201	51%

[a] Relative to control in parentheses

To determine optimal parameters for EMS seed-treatment, pre-germinated G254⊗ seeds were incubated for 16 h with varying EMS concentrations in the range 0.25-0.90% EMS. A pronounced germination delay was observed in the treated seeds and a linear correlation between M_1-seed lethality and EMS concentration was observed (Figure 2).

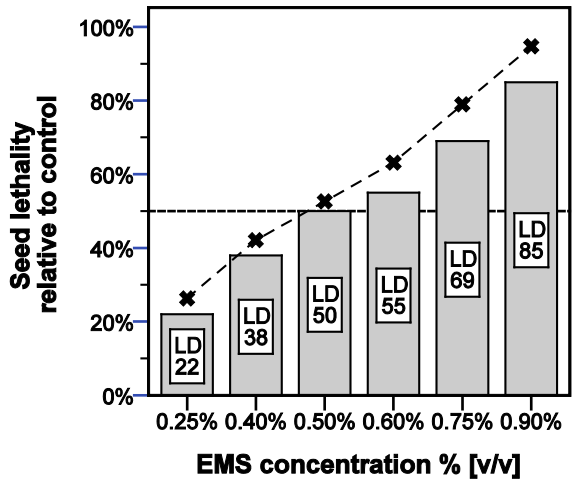

FIGURE 2. Seed lethality of seeds treated with several EMS concentrations. Bars represent the seed lethality at the different EMS concentrations. For each concentration the lethal dose (LD = lethality) is given. EMS concentrations are also plotted as black cross-markers (at 100× the concentration used), connected by an interrupted line.

A concentration of 0.6% EMS (LD_{55}) was used to treat a bulk of approximately 10,000 G254⊗ seeds to establishe a M_1 population of 1,824 EMS-mutagenized plants. Approximately half of these plants flowered and produced tubers in later stages of development.

Screening for mutations in candidate genes

Fifteen mutation-screening amplicons for eight different candidate genes involved in potato tuber quality traits were designed and used for HRM screening in either the pollen-treated or the seed-treated EMS populations (Table 2).

TABLE 2. Candidate genes used for mutation screening.

Gene	*Name*	*Chr.*	*Gene length (CDS)*	*Gene ID*	*Candidate gene reference*	*Screened populations (# of amplicons screened)*
LCY-e	Lycopene Epsilon Cyclase	12	6.7 kb (1.6 kb)	PGSC0003DMG400000333	Diretto et al. 2006	Pollen (1)
ZEP	Zeaxanthin Epoxydase	2	6.1 kb (2.0 kb)	PGSC0003DMG400004020	Römer et al. 2002	Pollen (2)
PWD	Phosphoglucan Water Dikinase	9	4.0 kb (1.9 kb)	PGSC0003DMG400016613	Kötting et al. 2005	Pollen (2) and seed (1)
GWD	Glucan Water Dikinase	5	15.5 kb (4.4 kb)	PGSC0003DMG400007677	Lorberthet al. 1998	Pollen (1) and seed (3)
PAIN	Potato Acid Invertase	3	19 kb (2.6 kb)	PGSC0003DMG400013856	Bhaskar et al. 2010	Seed (2)
SBE1	Starch Branching Enzyme I	4	7.6 kb (2.7 kb)	PGSC0003DMG400009981	Safford et al. 1998	Seed (1)
SBE2	Starch Branching Enzyme II	9	11 kb (3.6 kb)	PGSC0003DMG200002712	Jobling et al. 1999	Seed (3)
SSS3	Soluble Starch Synthase III	2	15 kb (4.2 kb)	PGSC0003DMG400016481	Abel et al. 1996	Seed (1)

Three amplicons had natural polymorphisms between the two parental G254 alleles. For these amplicons, we tested HRM mutation screening of both individual and pooled DNA (four plants per pool). Detection of differences in melting pattern between amplicons with novel mutations and the normal segregating melting patterns was problematic using pooled samples, but not using individual samples (Figure 3).

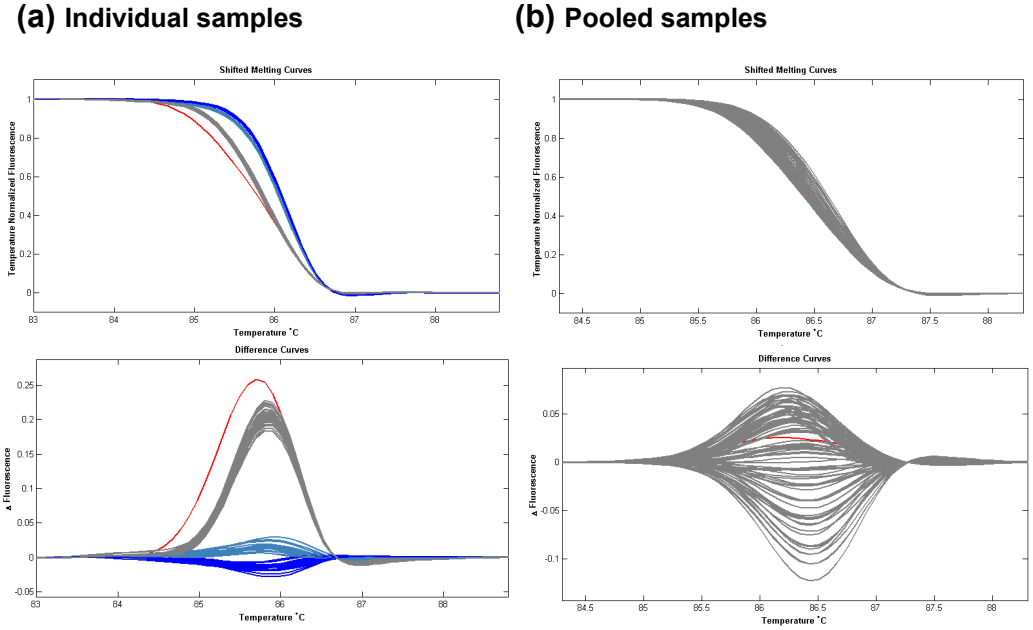

FIGURE 3. Example of HRM mutations screening (amplicon SBE2#1920) in individual and 4× pooled samples with an additional parental SNP. (A) Individual M1 plants segregate in three classes for the parental alleles; two homozygous classes (blue and light blue) and a heterozygous class (grey). Novel mutations have a distinct heterozygous melting pattern (red line). (B) Pools of four plants can have parental allelic copy-numbers varying from zero to eight and classification of the genotypes is not possible. Detection of a novel mutation (red line) in these pools is problematic.

The pollen-treated EMS populations were screened for six amplicons of four candidate genes, all located on different chromosomes (Table 2). Three amplicons were screened individually and three amplicons were screened using pooled DNA (four plants per pool) of all plants (Table S2). On average, 88% of the individually screened plants produced a good PCR product suitable for HRM screening. In total we screened 2,717 kb of the four candidate genes (Table 3). We detected and confirmed only one mutation, a C to T transition mutation found in an intron of the ZEP candidate gene (Table S4). The mutation was found in the 0.625% EMS pollen-treated population.

TABLE 3. Summary of mutation detection in candidate genes of the pollen-treated EMS populations.

EMS concentration	*Population size (% of total)*	*Screened target*[a]	*Verified mutations (intron mutations)*
0.625%	414 plants (22%)	597 kb	1 (1)
1.000%	1,269 plants (67%)	1,830 kb	0 (0)
1.500%	201 plants(11%)	290 kb	0 (0)
Total	**1,884 plants (100%)**	**2,717 kb**	**1 (1)**

[a] Calculated as target length (amplicon length - length of primers) × number. of screened plants. In case of pooled plants the number of screened plants was multiplied by the percentage of successfully screened plants in the non-pooled plants.

For the seed-treated EMS population leaf DNA of all 1,824 M_1 seedlings was extracted and screened for mutations. Six candidate genes, located on five different chromosomes (Table 2), were screened for eleven amplicons. Three amplicons were screened individually for all plants, six amplicons were completely screened using pooled DNA and two amplicons were individually screened for only part of the population (Table S3). On average 97% of the individual screened plants produced a good PCR product suitable for HRM screening. We screened over four million DNA base pairs of the seed-treated population and confirmed 65 mutations by sequencing the corresponding amplicons (Table 4). Of the 45 non-synonymous amino acid changes, 21 were predicted to have a conceivable impact on protein function. These included five stop codons in four different candidate genes. All but one mutation, a G to T transversion, were C-to-T or G-to-A transitions (Table S4). The overall mutation detection frequency was 1 mutation per 65.3 kb of screened bases. Mutation detection frequency of individually screened plants, pooled plants and individually screened plants with natural polymorphisms between the G254 alleles were comparable and in the range of 1 mutation/22-220 kb (Table S3).

TABLE 4. Summary of mutation detection in candidate genes of the seed-treated EMS population.

Gene	*Screened amplicons*	*Total target size (% CDS)*	*Screened target* [a]	*Verified mutations (intron mutations)*	*Mutation detection density*	*Non-synonymous mutations*	*Putative damaging mutations*	*Stop codon mutations*
GWD	3	798 bp (78%)	1,035 kb	20 (1)	1/52 kb	14	6	2
PAIN	2	474 bp (100%)	842 kb	16 (0)	1/53 kb	12	4	1
SBE2	3	645 bp (78%)	759 kb	4 (0)	1/190 kb	4	1	0
PWD	1	331 bp (63%)	588 kb	9 (3)	1/65 kb	6	4	1
SBE1	1	288 bp (100%)	512 kb	12 (0)	1/43 kb	6	3	0
SSS3	1	289 bp (100%)	510 kb	4 (0)	1/128 kb	3	3	1
Total	**11**	**2,825 bp (84%)**	**4,244 kb**	**65 (4)**	**1/65 kb**	**45**	**21**	**5**

[a] Calculated as target length (amplicon length - length of primers) × number of screened plants. In case of pooled plants the number of screened plants was multiplied by the percentage of successfully screened plants in the non-pooled plants.

Development of genetically stable M_2 and M_3 mutants

Using *in vitro* culture we tried to propagate twenty-six M_1 mutants, containing 27 of the most interesting mutations identified in the M_1 seedlings (one seedling contained mutations in two of the screened genes). Ten plants were dead by the time we started the *in vitro* propagation. Seventeen potentially chimeric mutant plants of the M_1 seed-treated EMS population, containing 18 putatively interesting mutations, were successfully propagated (Table 5). For

each successfully propagated mutant plant an average of eight independent *in vitro* clones were made. Of the 139 *in vitro* clones, 97 contained the identified mutation. For eleven mutants all clones contained the mutation, two mutations were lost in all clones, and for four mutants only a number of the *in vitro* clones contained the mutation. The vegetatively propagated clones originating from the mutant harbouring two of the identified mutations co-segregated for these mutations, with two of the clones having both mutations and the other nine clones having neither of the mutations. To obtain a genetically stable M_2 generation of the mutants, 39 *in vitro* clones containing 16 different mutations (1-4 plants per mutation) were used for selfing or for crossing with non-mutagenized RH. Selfing the mutant plants failed and crosses with the mutants as pollen donor were not attempted. Crosses with RH failed repetitively for six of the mutations, and included the clones harbouring the double mutation. For 10 of the 27 initially selected mutations, M_2 (M_1×RH) seeds were obtained (Table 5).

TABLE 5. Mutants selected for crossing with RH to obtain genetically stable M_2 and M_3 families.

Mutant	*Amino acid change*	*In vitro clones (M_1) (Mutant:Wildtype)*	*M_2 (M_1×RH) seeds*	*Segregation in M_2 (Mutant:Wildtype)*	*M_3 (M_2×M_2) seeds*
PAIN-13H02	T248I	8:0	191	38:11	Yes
PAIN-16H01	W299*	17:0	475	17:26	Yes
PWD-02H10	W579*	7:0	635	24:24	Yes
SBE1-06G12	G447R	5:0	621	1:64	*In progress*
GWD-10B07	W306*	10:0	18	*In progress*	*In progress*
GWD-11E07	R344W	2:0	548	*In progress*	*In progress*
SBE1-03C04	G440E	6:0	13	*In progress*	*In progress*
SBE2-11A10	D161N	10:0	582	*In progress*	*In progress*
SSS3-17F04	W392*	4:1	24	*In progress*	*In progress*
SSS3-05F10	A370V	11:0	225	*In progress*	*In progress*
GWD-16G10	E335*	3:0	No M_2 seed	No M_2 seed	No M_2 seed
PWD-07F12	S612L	8:4	No M_2 seed	No M_2 seed	No M_2 seed
SBE2-01C12	G413E	1:0	No M_2 seed	No M_2 seed	No M_2 seed
SBE2-04H02	G127R	1:4	No M_2 seed	No M_2 seed	No M_2 seed
PAIN/GWD-15A01[a]	D245N/H287Y	2:9	No M_2 seed	No M_2 seed	No M_2 seed
PWD-03H10	E602K	0:15	No *in vitro* mutants	No *in vitro* mutants	No *in vitro* mutants
SBE1-16B10	T432I	0:11	No *in vitro* mutants	No *in vitro* mutants	No *in vitro* mutants
GWD-05E11[b]	E322K	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
GWD-12E03[b]	E397K	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
GWD-15H11[b]	G350E	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
GWD-01B11[b]	G262S	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
PWD-16D03[b]	L610F	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
SBE1-02A10[b]	S433F	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
SBE1-09C08[b]	S454N	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
PAIN-17B03[b]	G270D	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants
SSS3-09G04[b]	L388F	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants	No *in vitro* plants

[a] Double mutant.
[b] Dead by the time *in vitro* propagation started.
* = stop codon

For four of the mutations M_2 seedlings were grown in 2011 and tested for presence of the mutations. Only one M_2 family showed the expected 1:1 segregation. Offspring of the other three crosses showed a (severely) distorted segregation (Table 5). Even so, M_2 plants containing the mutation were obtained for all four mutations. The single M_2 plant containing mutant SBE1-6G12 was too weak for immediate use in crossing towards an M_3 population. For the other three mutants, eight M_2 plants containing the mutation were crossed among each other and a large number of M_3 ($M_2 \times M_2$) seeds for genotypic and phenotypic evaluation were obtained.

DISCUSSION

Mutation detection using HRM

High resolution melting analysis is used as a sensitive and high-throughput mutation screening method in plants, animals and humans (BOTTICELLA *et al.* 2011; BUSH and KRYSAN 2010; GADY *et al.* 2009; ISHIKAWA *et al.* 2010; REED and WITTWER 2004). For species like potato, with high levels of nucleotide diversity between alleles (see previous Chapters), it is not always possible to find suitable candidate regions free of natural polymorphism between the two parental alleles. We show that HRM mutation screening for amplicons with parental SNPs is possible when the samples are screened individually. The mutation detection frequency in this case is comparable to individually screened amplicons without additional polymorphisms. Pooling of amplicons with parental SNPs is however not feasible, and it still has to be investigated how many additional SNPs and/or indels can be tolerated in individual amplicon sequencing. For amplicons without parental SNPs, we successfully screened DNA of both individual and 4× pooled plants for amplicons with a size range of 164 to 373 bp. The study of (GADY *et al.* 2009) shows a 4× pooling strategy is sufficiently sensitive for HRM detection of most mutations, while a 8× pooling is not. As we did not observe a difference in the mutation detection frequency of individually screened plants and pooled plants we confirm the high sensitivity of the HRM method using 4× pooled samples.

Mutations in the pollen-treated EMS populations

Pollen EMS treatment coupled to reverse genetic screening has previously been applied to maize (TILL *et al.* 2004). The EMS treatment method used there as well as in a study of EMS pollen-treatment in Eucalyptus (MCMANUS *et al.* 2006) is derived from Neuffer and Coe (1978) and is based on mixing paraffin oil-diluted EMS with pollen. In potato mass pollination is not possible and since EMS is a highly volatile, unstable compound hardly dissolvable in oil, the precise application of a standard dose to a single potato flower stigma is problematic. We therefore developed an alternative protocol in which EMS is dissolved in a sucrose solution of suitable osmolarity. The examination of pollen growth *in vitro* following exposure to EMS is frequently used to determine a suitable dose for pollen treatment (MCMANUS *et al.* 2007; NEUFFER and COE 1978). As we saw an effect of *in vitro* pollen germination on all tested EMS doses (0.5-2% EMS), the protocol we used should induce mutations in pollen. Furthermore, since most mutations are deleterious (CHARLESWORTLI and CHARLESWORTH 1998) it is expected that EMS treatment of pollen should *in vivo* be associated with a drop in embryo survival, fruit

retention, seed viability and seedling vitality (McManus *et al.* 2006). In contrast to the pollen-treatment study in Eucalyptus by McManus et al. (2006), clear effects for all these variables were noted in our study. Pollination success, number of seeds per berry, seed germination and the number of tuber-forming plants all decreased by increasing EMS concentrations. Pollination success and the number of seeds per berry, reflecting pollen fitness, dropped however much faster than the seed germination rate (vitality) and the percentage of tuber-forming plants (viability). Despite the sharp reduction in pollen fitness, the seeds that developed were thus relatively vital. Mutations reducing fitness in pollen are indeed unlikely to strongly correlate with reduced fitness in seedlings, as mutations reducing fitness in pollen are not inherited (McManus *et al.* 2006). Selection of most vital pollen in the styles of flowers therefore seems to have lowered the mutation density of the pollen-treated populations.

In total we screened over two-and-a-half million DNA base pairs of the pollen-treated populations and detected only a single mutation. The spontaneous mutation rate in potato is unknown, but given the spontaneous mutation rate in Arabidopsis of 10^{-7} to 10^{-8} bp/generation (Kovalchuk *et al.* 2000), it is not likely that the identified transversion mutation is a spontaneous mutation. Furthermore the mutation is a novel C to T transition, as expected from the alkylation of guanine by EMS (Sega 1984). We therefore conclude it is a genuine EMS-induced mutation. As only one mutation was found, no dependable estimation of the mutation frequency of the pollen-treated populations can be made. It appears however to be low. In the EMS pollen-treatment study in maize by Till et al. (2004) a mutation density of 1 mutation/500 kb was found. The mutation we found in the 0.625% EMS pollen-treated population was found after screening 597 kb of mutagenized sequence. However, for the 1% EMS pollen-treated population no mutations were found after screening 1,830 kb of mutagenized sequence. Continued screening of the 1.5% EMS treated population might have yielded additional mutations. However, since no mutations were found after screening 290 kb of sequence in this population, even in an optimistic scenario it is not expected that the mutation density per mutagenized genome will be more than half of the seed-treated population and thus, since in pollen-treated plants only one genome is mutagenized, will require at least four times more M_1 plants to be screened. Given the difficulties to obtain enough 1.5% EMS pollen-treated seeds, combined with (1) the high efficiency of screening seed-treated M_1 plants, (2) the high mutation density of the seed-treated population and (3) the high recovery of genetically stable M_2 mutants from the seed-treated population, we find mutagenesis by EMS seed-treatment and the subsequent screening of the chimeric M_1 population much more effective for potato.

Mutations in the seed-treated EMS population

Plants directly derived from mutagenized seeds are chimeric and thus a M_2 generation is most commonly used for mutant screening (Comai and Henikoff 2006; IAEA 1977). However, in diploid potato gametophytic incompatibility systems are active (Eijlander *et al.* 1997). This causes difficulties in the selfing of a mutagenized M_1 generation. Therefore, we used the self-compatible G254 genotype. Selfing a non-treated M_1 generation (I_1) of G254⊗, typically yields an I_2 generation for approximately 82% of the plants (Olsder and Hermsen 1976). We did

however not succeed in selfing the G254⊗ seed-treated M_1 EMS population and therefore directly screened the M_1 plants. In total we screened over four million DNA base pairs of the M_1 population and detected 65 melting curve alterations that were confirmed to be mutations by sequencing.

As expected for EMS-induced mutations, all but one of the mutations we detected were C-to-T or G-to-A transitions (98.5% transitions). A large number of EMS transversion mutations have been identified in rice, barley and tomato (CALDWELL *et al.* 2004; MINOIA *et al.* 2010; TILL *et al.* 2007) and few in Arabidopsis, maize and wheat (GREENE *et al.* 2003; SLADE *et al.* 2005; TILL *et al.* 2004). Therefore, the EMS mutagenesis response in potato seems similar to Arabidopsis, maize and wheat, with EMS-induced transversion mutations occurring at a low frequency. In the Arabidopsis study, it was proposed that the few transversion mutations could be either contaminants or spontaneous mutations (GREENE *et al.* 2003). We exclude that the transversion mutation we found is due to contamination. In earlier studies (Chapters 3 and 4), we extensively re-sequenced the potato genepool for the *StGWD* gene, the gene in which the transversion was detected and did not encounter the mutation. The novel mutation causes a pre-mature stop codon in the *StGWD* gene. Although no reproductive selective pressure has been applied to the mutant M_1 plant, we consider the chance of finding a spontaneous (loss-of-function) mutation in a gene of interest after screening approximately 1,800 potato plants low. In a study of *Solanum verrucosum* for example, only two pollen microspores among the 2.5 million microspores analyzed showed the spontaneous loss-of-function characteristics they were looking for (DE NETTANCOURT and DIJSTRA 1969). And in a study of Röntgen-irradiated monoploid potato (HOVENKAMP-HERMELINK *et al.* 1987) 12,000 mutagenized mini-tubers had to be screened to identify a loss-of-function mutation in the *GBSS* candidate gene.

Mutation density in comparison to other reverse genetics studies

The average mutation detection frequency of the seed-treated M_1 population was 1 mutation/65 kb of screened amplicon. The estimated mutation rate is thus 15.3 nucleotide substitutions for every 1 kb length screened in 1,000 M_1 plants. This mutation rate is higher than that of EMS-induced mutant populations of e.g. Arabidopsis (1/89-170 kb), rice (1/265-502 kb), tomato (1/332-737 kb) and melon (1/573 kb) (see Kurowska et al. (2011) for a summary and references to mutation rates in recent reverse genetics screens). When considering the number of mutagenized genomes, i.e. two genomes in diploid seeds, our result of 1 mutation/131 kb of mutagenized genome is in the same range as the mutation densities obtained for tetraploid and hexaploid wheat (1/150-240 kb of mutagenized genome) (BOTTICELLA *et al.* 2011; SLADE *et al.* 2005) and hexaploid oat (1/198 kb of mutagenized genome) (CHAWADE *et al.* 2010). In fact, the mutation density of our M_1 population is nearly equal to the highest mutation frequency obtained in a published EMS-mutagenesis screen so far, that of *Brassica rapa* (approximately 1/60 kb, or 1/120 kb of mutagenized genome) (STEPHENSON *et al.* 2010). All published reverse genetics mutant populations – except the pollen-treated EMS study in maize (TILL *et al.* 2004) and a seed-treated EMS study in diploid potato (MUTH *et al.* 2008) – are however screened at the M_2 generation.

Similar to our study, Muth et al. (2008) screened the M_1 generation of diploid potato. Seeds were treated with 1% EMS. Using Sanger sequencing, a series of 19 EMS-induced mutations in the *GBSS* gene were identified. The mutation density of 1/91 kb of this single-locus study is in the same range as the multi-locus results we obtained using 0.6% EMS. Although we used a much lower EMS concentration, the M_1 germination rate of both studies is similar. A noticeable difference between the two studies is that we pre-germinated seeds before treatment while Muth et al. (2008) treated dormant seeds. Cell-division cycle phases are thought to represent different levels of overall EMS-sensitivity. The S-phase was inferred to be the most sensitive one in EMS-treated barley seeds and reached within 24 hours of pre-soaking (NATARAJAN and SHIVASANKAR 1965). The comparable mutation detection frequency and germination rate of our study compared to the study of Muth et al. (2008), at a lower EMS concentration, thus indicates it might be beneficial to pre-germinate potato seeds before EMS treatment. A direct comparison between the different genotypes on mutation sensitivity was however not made. Therefore, the increase in mutation frequency could also be attributed to a genetic effect, or to small differences in the treatment protocol like the seed to EMS-solution ratio (IAEA 1977).

"Accessible" M_2 mutations

We sought to obtain genetically stable M_2 mutants for 27 of the 65 mutations we identified in the seed-treated EMS population. Since we screened the M_1 population at the seedlings stage we included weak, non-flowering and non-tuber forming plants. Of the 27selected mutations, 33% were found in non-vital plants that had already died by the time we started *in vitro* propagation. Of the 18 mutations in the more vital plants, we lost only two mutations (10%) in the *in vitro* M_1 clones due to chimeric tissue. Combined with a M_1-fertility rate of 62.5%, we obtained genetically stable M_2 mutant plants for 10 (56%) of these mutations. Of all 27 identified mutations for which we sought to obtain M_2 mutants we succeeded for 37% of the mutations. Even if only between one-third and a half of all mutations identified in the M_1 population are transformed into genetically stable M_2 mutants, one can still expect approximately 6-9 genetically stable "accessible" M_2 mutations for every 1 kb length screened in 1,000 M_1 plants (1 "accessible" mutation/118-176 kb). This is higher than the typical mutation density range of 1 mutation/200-500 kb found commonly in M_2 screening studies (KUROWSKA *et al.* 2011). Furthermore, the effort to obtain genetically stable M_2 mutations in our study could have been increased by screening at a later developmental stage of the M_1 population and by increasing the M_1-fertility rates. In vegetatively propagated species like potato, screening can be delayed until plants are mature and estimates on fertility can be made. Screening for example only M_1 plants that flower will increase M_1-fertility to a great extent while it is not expected to lead to a loss in mutant yield (HILDERING and VAN DER VEEN 1966). Furthermore, to select more vital plants and to reduce chimeric tissue, screening could have been commenced in the next vegetative generation of plants, formed by the tubers.

PERSPECTIVE – TOWARDS HOMOZYGOUS MUTANTS OF THE NOVEL ALLELES

Assumptions on the impact of amino acids changes of the identified mutant alleles needs to be proven by phenotypic tests. Mutations segregating in the M_2 (M_1×RH) generation are heterozygous and thus unless dominant unable to affect phenotype. To obtain homozygous mutants with potential phenotypes we are currently selecting and crossing M_2 (M_1×RH) plants harbouring the novel mutations and selecting M_3 plants homozygous for the mutations. By using a strategy of crossing different M_2 family plants among each other, we reduce the number of undesirable background mutations in the M_3 generation. To integrate the novel alleles into elite tetraploid breeding material for crop improvement, additional crosses have to be made that will further reduce the number of potentially harmful background mutations.

Focus of our study were candidate genes involved in sugar and starch metabolism. Phenotypic effects associated with these candidate genes have been shown in knockout and silencing studies (Table 2). We identified five premature stop codons in four of these candidate genes. The stop codon mutations are likely to create loss-of-function phenotypes similar to the earlier studies, without the use of genetic modification. One of these mutations was lost in the M_1 generation, but for the other four we have obtained a M_2 generation and for two of these we already obtained homozygous M_3 plants that are being evaluated at the molecular and phenotypic level. For the two other loss-of-function mutations and for most of the other mutant alleles we are selecting M_2 plants with the mutations that can be used for crossing into an M_3 generation.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Nick de Vetten of Averis Seeds, Valthermond, The Netherlands, and Ir. Guus Heselmans of Meijer Potato, Rilland Bad, The Netherlands for assistance in growing the seed-mutagenized EMS population and crossing identified mutants. We thank Isolde Pereira for maintaining the *in vitro* plants and Dirk Jan Huigen for crossing the M_2 plants. We thank Ir. Peter Vos for assistance with HRM screening. This research was supported by a grant of the Dutch technology foundation STW, project WPB-7926.

SUPPORTING INFORMATION

TABLE S1. Primer sequences of candidate genes used for mutation screening.

Gene	Amplicon	Chr.	Target size[a] [bp]	Screened EMS Population	Forward primer	Reverse Primer	T_a (°C)
LCY	LCY#7	12	273	Pollen	AAAAACACTTGCATTTGGTG	TTTGGAGCTTCAGACAGTGA	57
ZEP	ZEP#3	2	320	Pollen	AGCCAAGAAAACCTGCTTTA	CACCAACTCTTCCAGGATGT	57
ZEP	ZEP#6	2	263	Pollen	GTTACATTGCTTGGGGACTC	GATGATATCCACAGGGCTTC	57
PWD	PWD#5	9	174	Pollen	TCTTAGTTTCTGTAGCTAAAATGTCA	AAGCTAAACGCAAAAATGAAG	56
PWD	PWD#1	9	331	Pollen & Seed	AGCATCAGTTGCATATGGTT	CCCAAAAATTTTCAGGATTG	60
GWD	GWD#2	5	281	Pollen & Seed	GCGCATAAATCTGGGTATTC	TTTAGATAAGGCCGGTGGT	59
GWD	GWD#2728	5	286	Seed	CCTCCACTGCTGTTAGACACTT	CGCACACTACATAAACTGTAAG	57
GWD	GWD#2930	5	231	Seed	GTGAGTGTCAAGCATTCCGATT	ACTAGATACCCCTGTCTATCAA	57
PAIN	PAIN#0607	3	330	Seed	GGTCAAGTACAAAGGCAACCC	AAGTCATACGTCCCAATAGCA	57
PAIN	PAIN#0910	3	144	Seed	ATAACAGAGTATTCCAAGGACA	ACAAAGTCGCGCAACAAACCTC	57
SBE1	SBE1#0304	4	288	Seed	CCTCAATGGCTTTGATATTGG	TCTGGGAAAATCTTGTGAATC	56
SBE2	SBE2#1516	9	122	Seed	AAACACAGTTGCAATAGTAAA	TTGTCGCAGTTATCATGTCAC	56
SBE2	SBE2#1920	9	249	Seed	ATTTTATGTTTTAGGCATCAAC	GTTTGTCAAAAGGGGGTCTATT	56
SBE2	SBE2#2122	9	274	Seed	CGCTTGCTCTCAGGATTTGCTC	AGTTCTCCTTCAAGTGGCATTC	56
SSS3	SSS3#F2R3	2	289	Seed	AGATTATTTTGTTTTCCAGAGG	CTCAAGCAAGAAATTTTCAAAG	56

[a] Amplicon length - length of the primers

TABLE S2. Per amplicon summary of mutant detection in the pollen-treated EMS populations.

Gene	*Amplicon*	*Target size*	*CDS target size [bp]*	*SNPs in G254 alleles*	*Screen method*	*Screened plants*	*Total bp screened*[a]	*Verified mutations*
GWD	GWD#2	281 bp	233	No	Individual	1,628	457 kb	0
PWD	PWD#5	174 bp	42	No	Individual	1,740	305 kb	0
PWD	PWD#1	331 bp	209	No	Pooled	1,708	565 kb	0
ZEP	ZEP#6	263 bp	110	No	Individual	1,597	420 kb	1
ZEP	ZEP#3	320 bp	138	No	Pooled	1,748	559 kb	0
LCY	LCY#7	273 bp	45	No	Pooled	1,504	411 kb	0
						Total	**2,717 kb**	**1**

[a] Calculated as target size × number of screened plants. In case of pooled plants the number of screened plants was multiplied by the percentage of successfully screened plants in the non-pooled plants.

TABLE S3. Per amplicon summary of mutant detection in the seed-treated EMS population.

Gene	*Amplicon*	*Target size*	*CDS target size [bp]*	*SNPs in G254 alleles*	*Screen method*	*Screened plants*	*Total bp screened*[a]	*Verified mutations*	*Mutation detection density*	*Putative damaging mutations*	*Stop codon mutations*
GWD	GWD#2728	286 bp	225	No	Pooled	All pools	508 kb	8	1/63.5 kb	2	1
GWD	GWD#2	281 bp	233	No	Individual	1796	505 kb	11	1/45.9 kb	2	1
GWD	GWD#2930	231 bp	163	No	Individual	96	22 kb	1	1/22.2 kb	0	0
PAIN	PAIN#0607	330 bp	330	No	Pooled	All pools	586 kb	14	1/41.9 kb	3	1
PAIN	PAIN#0910	144 bp	144	No	Pooled	All pools	256 kb	2	1/127.9 kb	0	0
PWD	PWD#1	331 bp	209	No	Pooled	All pools	588 kb	9	1/65.3 kb	2	1
SBE1	SBE1#0304	288 bp	288	No	Pooled	All pools	512 kb	12	1/42.6 kb	3	0
SBE2	SBE2#1920	249 bp	249	1 SNP	Individual	1770	441 kb	3	1/146.9 kb	0	0
SBE2	SBE2#1516	122 bp	108	No	Pooled	All pools	217 kb	1	1/216.7 kb	1	0
SBE2	SBE2#2122	274 bp	145	1 SNP, 1 InDel	Individual	369	101 kb	0	NA	0	0
SSS3	SSS3#F2R3	289 bp	289	1 SNP	Individual	1763	510 kb	4	1/127.4 kb	2	1
						Total	**4,244 kb**	**65**	**1/65.3 kb**	**15**	**5**

[a] Calculated as target size × number of screened plants. In case of pooled plants the number of screened plants was multiplied by the percentage of successfully screened plants in the non-pooled plants.

TABLE S4. List of all identified mutants. The first mutation (underlined) was found in the pollen-treated population. All other mutations were found in the seed-treated population.

Mutant name	*Gene amplicon*	*Type and position of the mutation*	*Amino acid change*	*Non-synonymous mutation*	*Prediction on protein effect*	*In vitro plants*[a]
ZEP-18C07	ZEP#6	8277 C/T	intron	No	Tolerated	NA
GWD-03A03	GWD#2	8176 G/A	V395I	Yes	Tolerated	NA
GWD-05E11	GWD#2	7957 G/A	E322K	Yes	Damaging	0
GWD-08B04	GWD#2	8011 C/T	L340F	Yes	Tolerated	NA
GWD-10A09	GWD#2	7998 G/A	E335E	No	Tolerated	NA
GWD-11E07	GWD#2	8023 C/T	R344W	Yes	Tolerated	2
GWD-12E03	GWD#2	8182 G/A	E397K	Yes	Tolerated	0
GWD-13C12	GWD#2	7983 G/A	L330L	No	Tolerated	NA
GWD-13E12	GWD#2	8185 G/A	E398K	Yes	Tolerated	NA
GWD-15H11	GWD#2	8042 G/A	G350E	Yes	Damaging	0
GWD-16G10	GWD#2	7996 G/T	E335*	Yes	Damaging	3
GWD-19C01	GWD#2	7922 G/A	intron	No	Tolerated	NA
GWD-16G02	GWD#2728	7635 G/A	E254K	Yes	Tolerated	NA
GWD-01B11	GWD#2728	7659 G/A	G262S	Yes	Damaging	0
GWD-19E10	GWD#2728	7676 C/T	D267D	No	Tolerated	NA
GWD-02G09	GWD#2728	7689 C/T	L272L	No	Tolerated	NA
GWD-15A01	GWD#2728	7734 C/T	H287Y	Yes	Damaging	11
GWD-02F09	GWD#2728	7747 G/A	S291N	Yes	Tolerated	NA
GWD-10B07	GWD#2728	7792 G/A	W306*	Yes	Damaging	10
GWD-12D12	GWD#2728	7811 G/A	P312P	No	Tolerated	NA
GWD-19B11	GWD#2930	10317 C/T	T643I	Yes	Tolerated	NA
PWD-10A12	PWD#1	7238 G/A	V558I	Yes	Tolerated	NA
PWD-14D12	PWD#1	7438 C/T	intron	No	Tolerated	NA
PWD-16D03	PWD#1	7394 C/T	L610F	Yes	Damaging	0
PWD-02B10	PWD#1	7460 C/T	intron	No	Tolerated	NA
PWD-02G02	PWD#1	7346 G/A	G594S	Yes	Tolerated	NA
PWD-02H10	PWD#1	7302 G/A	W579*	Yes	Damaging	7
PWD-02H11	PWD#1	7438 C/T	intron	No	Tolerated	NA
PWD-03H10	PWD#1	7370 G/A	E602K	Yes	Damaging	15
PWD-07F12	PWD#1	7401 C/T	S612L	Yes	Damaging	12
SBE1-02A10	SBE1#0304	4210 C/T	S433F	Yes	Damaging	0
SBE1-03C04	SBE1#0304	4231 G/A	G440E	Yes	Damaging	6
SBE1-04F03	SBE1#0304	4154 G/A	L414L	No	Tolerated	NA
SBE1-05G05	SBE1#0304	4091 G/A	L393L	No	Tolerated	NA
SBE1-06G12	SBE1#0304	4251 G/A	G447R	Yes	Damaging	5
SBE1-07C08	SBE1#0304	4256 C/T	N448N	No	Tolerated	NA

Mutant name	*Gene amplicon*	*Type and position of the mutation*	*Amino acid change*	*Non-synonymous mutation*	*Prediction on protein effect*	*In vitro plants*[a]
SBE1-07G05	SBE1#0304	4316 C/T	A468A	No	Tolerated	NA
SBE1-08G11	SBE1#0304	4148 C/T	S412S	No	Tolerated	NA
SBE1-09C08	SBE1#0304	4273 G/A	S454N	Yes	Possibly damaging	0
SBE1-16B10	SBE1#0304	4207 C/T	T432I	Yes	Damaging	11
SBE1-16G05	SBE1#0304	4091 G/A	L393L	No	Tolerated	NA
SBE1-19G01	SBE1#0304	4147 C/T	S412F	Yes	Damaging	NA
SBE2-01C12	SBE2#1516	11439 G/A	G413E	Yes	Damaging	1
SBE2-04H02	SBE2#1920	2988 G/A	G127R	Yes	Tolerated	5
SBE2-11A10	SBE2#1920	3090 G/A	D161N	Yes	Tolerated	10
SBE2-15E12	SBE2#1920	2989 G/A	G127E	Yes	Tolerated	NA
PAIN-05H01	PAIN#0607	2566 G/A	G314G	No	Tolerated	NA
PAIN-05H06	PAIN#0607	2489 G/A	V289M	Yes	Tolerated	NA
PAIN-09H09	PAIN#0607	2477 C/T	L285L	No	Tolerated	NA
PAIN-11G01	PAIN#0607	2383 C/T	P253P	No	Tolerated	NA
PAIN-13H02	PAIN#0607	2367 C/T	T248I	Yes	Damaging	8
PAIN-14B04	PAIN#0607	2498 G/A	A292T	Yes	Tolerated	NA
PAIN-14F02	PAIN#0607	2376 C/T	T251I	Yes	Tolerated	NA
PAIN-15A01	PAIN#0607	2357 G/A	D245N	Yes	Damaging	11
PAIN-16C06	PAIN#0607	2315 G/A	V231I	Yes	Tolerated	NA
PAIN-16H01	PAIN#0607	2520 G/A	W299*	Yes	Damaging	17
PAIN-16H11	PAIN#0607	2505 C/T	P294L	Yes	Possibly damaging	NA
PAIN-17B03	PAIN#0607	2433 G/A	G270D	Yes	Damaging	0
PAIN-18B11	PAIN#0607	2589 C/T	P322L	Yes	Tolerated	NA
PAIN-18F08	PAIN#0607	2315 G/A	V231I	Yes	Tolerated	NA
PAIN-18H01	PAIN#0910	3228 C/T	A463V	Yes	Possibly damaging	NA
PAIN-19F01	PAIN#0910	3206 C/T	L456L	No	Tolerated	NA
SSS3-09G04	SSS3#F2F3	2923 C/T	L388F	Yes	Possibly damaging	0
SSS3-05F10	SSS3#F2F3	2870 C/T	A370V	Yes	Damaging	11
SSS3-17F04	SSS3#F2F3	2937 G/A	W392*	Yes	Damaging	5
SSS3-10D05	SSS3#F2F3	2976 G/A	R405R	No	Tolerated	NA

[a] NA = not attempted
* = stop codon

CHAPTER 6

General Discussion

INTRODUCTION

The research in this thesis is focussed on sampling and creating genetic diversity related to genes involved in quality traits of potato tubers and processed products in the framework of the project "Potatoes with novel properties for consumption and processing industry". The aim of the research project proposal was "to identify unexploited genetic diversity, or to create novel mutant alleles, in genes involved in important properties of potato". To address these aims, this thesis studies the challenges of reliable, high-throughput identification and genotyping of sequence variants in existing potato tetraploid cultivar panels, and explores the efficiency of ethyl methanesulphonate (EMS) mutagenesis in combination with reverse genetics DNA screening in potato.

DISCOVERY AND GENOTYPING OF DNA SEQUENCE VARIANTS IN EXISTING GERMPLASM

Potato breeding germplasm is composed of tetraploid progenitor clones and cultivars and a small number of diploid genotypes. A comprehensive overview of all named cultivars is found in the pedigree database (>8000 records) (VAN BERLOO *et al.* 2007). To obtain insight into the potato breeding germplasm, we applied pedigree based network analysis, using close to 6,000 interrelated cultivars and progenitor lines present in the potato pedigree database. One striking observation is the interconnectivity of potato germplasm (Figure 1). Furthermore, the analysis clearly illustrates the clustering of potato cultivars by continent, caused by a deficiency of inter-continental crosses, and pinpoints major potato germplasm contributors and ancestral potato founding lines. Only a limited number of clones seem to have contributed to the current gene pool. These include: (1) land race type cultivars acting as founding fathers for modern breeds (2) major contributing ancestors which are cultivars that have parented large numbers of cultivars and (3) Latin American *Solanum* species that have been used to introgress pathogen resistance. The pedigree database and sources like the World Catalogue of Potato Varieties show that almost all cultivated germplasm of the past centuries (not including Latin American landraces) is still available in cultivars that are maintained today, and that allelic diversity can thus be sampled exhaustively.

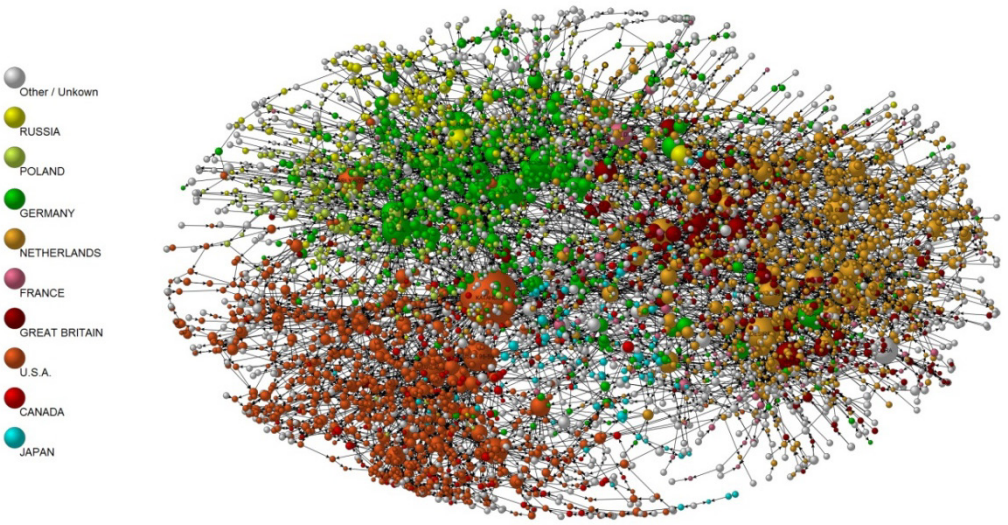

FIGURE 1. Potato pedigree analysis network. Vertices represent 5,991 related cultivars from the Potato Pedigree database. (http://www.plantbreeding.wur.nl/potatopedigree/). All direct parent → offspring relationships are shown. Size of the vertices is proportional to the number of offspring a cultivar has. Vertices are colour coded for the nine most represented countries of origin.

For the experiments in this thesis we used samples of the potato germplasm that were selected according to the concept of core collections (BROWN 1989). A core collection is defined as the smallest possible number of genotypes comprising the largest possible amount of genetic variation. Furthermore, the throughput of the different experiments resulted into a choice for different subsets of the potato genepool. In Chapters 2 and 3 we made use of a core set of 220 potato cultivar and progenitor lines and an extended set of 190 breeding lines selected by D'hoop et al. (2008). The core set represents worldwide commercial potato germplasm. From this core set the population structure was analysed previously and cultivars have been phenotyped for important agronomical and quality related traits for a number of subsequent growing seasons (D'HOOP *et al.* 2010; D'HOOP *et al.* 2008). In Chapter 4, we limited the set of cultivars to 84 because of the anticipated read depth requirements for quantitative genotyping. For this chapter, a representative sample of the core set was selected on basis of AFLP-based genetic distances.

Genes targeted for genotyping-by-sequencing (GBS)

A genotyping-by-sequencing (GBS) approach was applied in this thesis for discovery of sequence variants in existing potato breeding germplasm. In a GBS approach a DNA sample is sequenced using first, second or third-generation sequencing technology and sequence variants are called by comparing the resequenced fragments to a reference or consensus sequence. As they are discovered, sequence variants are characterized in terms of their reference (genome) sequence position and genotyped in individual samples. The discovered

variants and genotyping results can be directly used for genetic analysis (VARSHNEY *et al.* 2009).

Depending on the throughput, GBS can be equally broad and whole genome oriented, or targeted on specific candidate genes. In Chapters 2 and 3 of this thesis we used traditional Sanger DNA sequencing to screen four previously identified potato candidate genes in large populations with unrelated cultivars, as well as full sibs. In Chapter 2, we screened genes of the carotenoid biosynthesis pathway (*CHY2*, *LCYe*, *ZEP*) implicated in tuber flesh color, and in Chapter 3 we screened a candidate gene involved in starch phosphorylation (*StGWD*), which has been linked to cold-induced sweetening. In Chapter 4 we made use of second-generation, or next-generation, massively parallel sequencing (MPS) to scale up variant discovery and make it more whole genome-oriented, with over eight hundred genes re-sequenced and genotyped in over eighty existing cultivars.

The number of candidate genes for potato quality traits proposed in literature is limited, as for many important phenotypic traits the molecular basis is still poorly understood. For traits like tuber size distribution, eye depth, russet skin type, or adaptation (i.e. the stability of yield and quality across a wide gradient of agro-ecological environments), for example, the potato literature does not provide insight in the underlying processes, the underlying physiology or biochemistry, the underlying proteins, or underlying genes. For other traits, such as cooking type, tuber flesh colour or cold-induced sweetening it is obvious to propose genes from cell wall, carotenoid or carbohydrate metabolism respectively. Unfortunately, the number of candidate genes provided by the scientific literature limited us to ~100 candidate genes with clear relations to potato traits. The genes targeted in Chapter 4 were therefore expanded to represent a broader set of functional genes. Selection of this set was based on putative gene functions in both primary- and secondary metabolic pathways, potato quality traits and biotic and abiotic stresses. Furthermore, the set includes a large set of conserved orthologous sequence genes (COSII) useful for genetic anchoring and phylogenetic studies in broad germplasm (WU *et al.* 2006). In addition, the target sequences included a number of intergenic regions corresponding to putative AFLP marker sequences that previously showed genetic association with important potato quality traits for validation (D'HOOP *et al.* 2008), and a number of chloroplast and mitochondrial genes.

Mapping sequences to the potato reference genome

The Potato Genome Sequencing Consortium has recently published the genome sequence of the doubled monoploid 1-3 516R44 (DM) genotype derived from *Solanum tuberosum* group Phureja (XU *et al.* 2011). This DM reference genome has been sequenced to high accuracy and is assembled into large superscaffolds, and is thus expected to be used as the universal reference against which other sequences are aligned. In Chapter 4 we used a mapping approach for aligning resequencing data with the DM genome (XU *et al.* 2011) and identify sequence variants and genotypes based on the superscaffold coordinates of DM. An advantage of alignment to an annotated reference genome like DM is that it allows to predict whether a sequence variant falls within or near a gene of interest, and whether it is expected to cause a functional change in the protein product (synonymous vs. non-synonymous change)

that might alter the enzyme activity of the protein. This can be very useful in determining whether a particular sequence variant is likely to be responsible for a phenotype of interest.

A disadvantage of mapping sequence reads toward a reference sequence is that structural variation like chromosome rearrangements, inversions and large (transposon) insertions are likely to be missed. This disadvantage can be partly avoided by *de novo* assembly. However, computational difficulties associated with the assembly of highly diverse polyploid species like potato makes mapping sequence reads to a reference sequence a more straightforward approach. In the MPS approach of Chapter 4, we did not systematically look for structural variants, nor did we consider gene copy number variants (CNVs). Methods for detecting both structural variants and CNVs using a mapping approach have however been developed (KRUMM *et al.* 2012; LAM *et al.* 2010) and could be applied.

In Sanger sequencing amplicons are individually evaluated for length, and large insertions are more easily detected. Large transposon insertions like the one in allele 1of ZEP, causal to zeaxhantin accumulation and detected by Sanger sequencing in Chapter 2, are at this time not identified in the MPS results of Chapter 4, although present. Fortunately, sequence variants in perfect linkage disequilibrium (LD) with these causal transposons are frequently detected. In fact, we initially detected the marker-trait association of the ZEP allele through assaying other sequencing variants, in LD with the transposon insert.

Quantitative genotyping of tetraploids using GBS

The fundamental prerequisite for the assignment of (multi-allelic) allele configurations in a polyploid species with polysomic inheritance like potato is the accurate reproduction of allele copy number. In Chapters 2 and 3 we show that, in line with previous studies (DE KOEYER *et al.* 2009; RICKERT *et al.* 2002; SATTARZADEH *et al.* 2006), Sanger amplicon sequencing can be used as a reliable GBS method to quantify allele copy number of discovered sequence variants. In Sanger amplicon sequencing, sequence variants in amplicons of target genes are identified in the sequence chromatograms and directly quantified. Allele copy number differences are recorded on the basis of peak height, or area, in sequence chromatograms. A beneficial feature of direct Sanger sequencing to avoid genotyping errors is the joint observation of the increase in peak height of one allele to the expense of a decrease in peak height of the alternative allele (WECKX *et al.* 2005).

The Sanger amplicon sequencing strategy, although labour intensive, has been successful when the goal is to identify and genotype sequence variants of a broad gene pool for a single or a limited number of target genes, like the four candidate genes resequenced in Chapters 2 and 3. To avoid unequal amplification efficiency of different alleles with Sanger amplicon sequencing, primers have to be designed in regions that do not contain (undetected) polymorphisms relative to the primer annealing sites, and PCR reactions need optimization. As sequence variants are more common in intron sequences, primers designed in Chapters 2 and 3 were based on the more conserved exon sequences. Furthermore, for each amplicon that is Sanger sequenced, single locus amplification had to be verified and indel polymorphisms avoided, since these can result in undecipherable sequence chromatograms. Given the

relatively short length of exons, Sanger amplicons however commonly include introns. In a genetically diverse species like potato, that is highly polymorphic and heterozygous at the DNA sequence level, indel polymorphisms occur on average once every 163 bp (Chapter 4). This results in many DNA sequences unsuitable for direct Sanger sequencing and makes the Sanger-based GBS method time consuming.

Second-generation MPS, with its high throughput, currently provides an effective GBS alternative to Sanger amplicon sequencing (ELSHIRE *et al.* 2011; NORDBORG and WEIGEL 2008). Similar to Sanger-based sequencing, MPS can be used to directly identify and genotype variants using a GBS approach, given that sequences are obtained from an adequately large representation of individuals with sufficient read depth and that the source of sequences can be tracked using an index (GRATTAPAGLIA *et al.* 2011; VARSHNEY *et al.* 2009). In Chapter 4 we designed simple index sequences fused to the sequencing adapters to create custom-indexed MPS libraries. The inclusion of these indices permits tracking of the alleles' cultivar source so that zygosity can be determined in individual genotypes, allowing GBS of many individuals. We developed indexes to multiplex 12 samples in a single sequencing lane and sequenced 84 samples in total. Larger numbers of indices can however be easily created (ELSHIRE *et al.* 2011) and allow multiplexing of tens to hundreds of individuals.

Complexity reduction

Complexity reduction limits the portion of genetic material to be resequenced to that considered relevant for the question at hand and is used to increase sequence depth for loci of interest. The portion of the genome that is sequenced can be reduced by applying sequence capture methods such as SureSelect, Nimblegen, and Raindance (GNIRKE *et al.* 2009; KISS *et al.* 2008; NIJMAN *et al.* 2010) (GNIRKE *et al.* 2009; KISS *et al.* 2008; NIJMAN *et al.* 2010), and methods such as Crops, RNA-Seq and Rad-seq (MAMANOVA *et al.* 2010). In this thesis we applied the SureSelect enrichment method to capture target loci of interest. The genes we targeted were mainly single-copy genes, selected to avoid potential difficulties in distinguishing allelic variants from paralogous sequences (NG *et al.* 2009).

An inconvenient aspect of the SureSelect method was the huge amount of enriched chloroplast sequences, while only a few baits targeting chloroplast sequences were included in the bait library. We explained this effect on the basis of the large ratio difference between chloroplast and nuclear genome copies per leaf cell. It severely reduced the overall sequencing depth of the nuclear targets. Fortunately, MPS output increases constantly and the costs of DNA sequencing correspondingly drops. We initially planned to sequence the 84 cultivars in a single Illumina GAII sequencing run in Chapter 4. During the course of the experiment the Illumina HiSeq with increased sequence output, became available. We used this platform for sequencing the libraries. Despite the large fraction of cpDNA, the remainder of genomic DNA sequences resulted in a median per cultivar average sequence depth of 63× (ranging from 15× to 177×) for the covered or "accessible" genome, and a sequence depth of 88× for the regions directly targeted by SureSelect enrichment.

Analysis of MPS data for quantitative genotyping

To estimate allele copy number of a sequence variant using MPS, the ratio between reads with a reference allele and reads with an alternative allele in an individual plant is used (Figure 1).

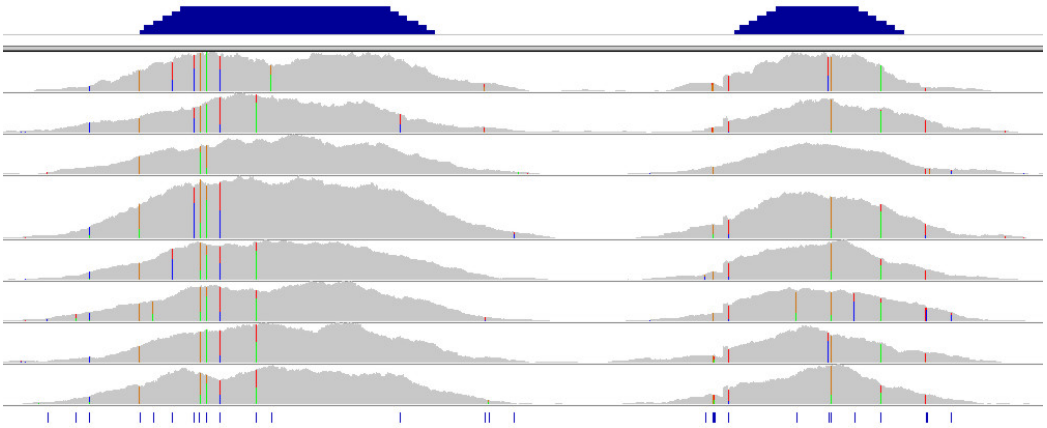

FIGURE 1. Genotyping-by-sequencing using next-generation sequencing in potato. The illustration shows a 2.4 kb part of the potato genome. At the top, in blue, the tiled (1 bait per 20 bp) 120-bp SureSelect baits are shown for two adjacent target regions of a gene. In the parts below, the sequence read depth (varying from 0 to ~100×) of eight potato samples are shown. The lowermost part of the illustration shows, in blue, all SNP positions in the accessible (i.e. sufficiently covered) part of the genome. At these positions, the number of reads with an A (green), T (red), C (blue) or G (brown) nucleotide are shown for samples not homozygous for the reference nucleotide (i.e. grey in samples without variation at this position). Based on the relative read depth of the SNP alleles, allele copy number (for a binary SNP: nulliplex, 0%; simplex, 25%; duplex, 50%; triplex, 75% and quadruplex 100% of non-reference nucleotides) of each SNP can be clearly identified, at least at the higher read depths.

Currently, most MPS genotype-calling algorithms are designed to call haploid and diploid genotypes, not polyploids. The genetic variant detector Freebayes (The MarthLab, https://github.com/ekg/freebayes) is an exception to this, suitable for genotyping autotetraploids as well as other polyploids. Like most MPS genotype callers, Freebayes provides a read-based likelihood estimate that a called genotype is wrong. This genotype quality (GQ) parameter is encoded as a phred quality score, $-10\times\log_{10}(p)$. In Chapter 4, we validated the Freebayes called genotypes for a small subset of 270 binary SNPs in an independent SNP assay. To reduce the number of conflicting genotype calls, we applied a GQ26 filter. The GQ estimated error rate at this threshold is 0.25%, whereas in practice we saw an overall error rate of around 2%, and around 4% for duplex calls, using this threshold. The GQ estimate assumes no ascertainment bias; e.g. the relative number of allele-specific sequencing reads is expected to be proportional to the zygosity. In a highly diverse species like potato, and by using hybridisation-based target enrichment, a bias that favours alleles more similar to the capture baits and/or genome reference sequence can however exist. We tried to minimize this allele-specific bias by tiling the 120-bp SureSelect capture baits, with one new probe starting approximately every 20 bp (tiling depth of 6×) and by using relaxed read

mapping quality requirements during genotype calling. The higher than expected error rate at GQ26 is probably a result of some remaining allele-specific read bias. Genotype calling algorithms should include the option to compensate for this skewed allelic bias. Application of a model-based clustering method that assigns clusters of samples with similar (underrepresented) read ratios to a single genotype class, as for example incorporated in fitTetra (VOORRIPS *et al.* 2011) could also be applied to MPS genotype calling. Furthermore, expectations on the genotype proportions according to the underlying allele frequency, assuming Hardy-Weinberg equilibrium (HWE), could be incorporated. Improved allele calling algorithms specially designed for polyploids and pooled analyses are being developed and we are looking forward to continued methodological and technical improvements in this field.

What sequence read depth is required for quantitative genotyping of tetraploids?

The reliability of a tetraploid genotype call depends on both read depth and zygosity class. At a read depth of approximately 15×, all four alleles of a tetraploid sample are expected to be sequenced at least once (GRIFFIN *et al.* 2011). It can thus be expected that the three classes of heterozygotes can be distinguished from homozygotes at this minimal depth, but to make a distinction within these three classes of heterozygotes (simplex, duplex and triplex), higher read depths are required. We initially estimated, that a sequence depth of at least 48× is required for this distinction (see Chapter General Introduction). In practice, we observed that a target sequence depth of around 80× was more appropriate for reliable quantitative genotyping.

An alternative to quantitative genotyping using GBS is to sequence target regions at low coverage per individual, accepting that a sequence variant can only be genotyped as a dominant presence/absence marker. In the past, the research community could accept dominantly scored molecular marker systems as well and in Chapter 4 we speculate that for marker-trait analysis in potato, dominant data will suffice. For dominant GBS marker analysis, the read depths required are ~4 to 5-fold lower than the 80x read depth proposed in this thesis as adequate for quantitative genotyping. Had we not aimed for quantitative genotyping in Chapter 4 and not lost so much sequence reads to chloroplast sequences, we expect that well over 500 potato cultivars could have been dominantly genotyped for the same 1.44 Mb target region using GBS. For this large number of samples –where each sample needs to be individually indexed – more efficient protocols for sequence library preparation and automation of the library preparation are required.

Comparison of Sanger- and MPS-based GBS results

The candidate genes sequenced in Chapters 2 and 3 were also included as MPS target in Chapter 4. Comparison of the sequence covered by both the Sanger and MPS method shows that both the number of discovered variants and the allele copy numbers of those variants matched very well between the two methods. The MPS method was however more fit to identify very rare alleles, for example, those occurring in only one sample, in simplex condition. The signal-to-noise ratio in Sanger sequencing complicates the discovery of these

kind of very rare sequence variants in tetraploid cultivars. In MPS sequencing, these cultivar-restricted, simplex variants are however clearly detected given that they are covered by an adequate number of sequence reads to discriminate them from sequencing errors (e.g. by at least 5 observations of the alternative sequence variant in a single cultivar).

DEVELOPMENT OF HIGH-THROUGHPUT SNP GENOTYPING ASSAYS

Currently, high-throughput SNP genotyping assays are commonly used to screen a large number of individuals for large numbers of markers (GANAL *et al.* 2011; LIJAVETZKY *et al.* 2007). Both high- and low density assays can be custom-designed and can provide the ability to interrogate tens to thousands of sequence variants, across hundreds to thousands of genotypes. Automated low-density SNP assays like KASP, TaqMan, Fluidigm and Golden Gate assays can be used for targeted analysis in applications like map-based cloning, cultivar identification and introgression breeding. For the more marker-dense applications like marker-trait association analysis, the density of SNP genotyping platforms such as Infinium or Affymetrix arrays is such that they can be expected to be in LD with QTL alleles for any trait of interest.

Similar to MPS- and Sanger-based GBS, high-throughput SNP genotyping platforms like KASP, and Infinium arrays have the potential to allow discrimination between the five classes of allele copy numbers of a tetraploid species (Figure 2). Recently, software for automated genotype calling of tetraploid species using data from these kinds of platforms has been developed (VOORRIPS *et al.* 2011).

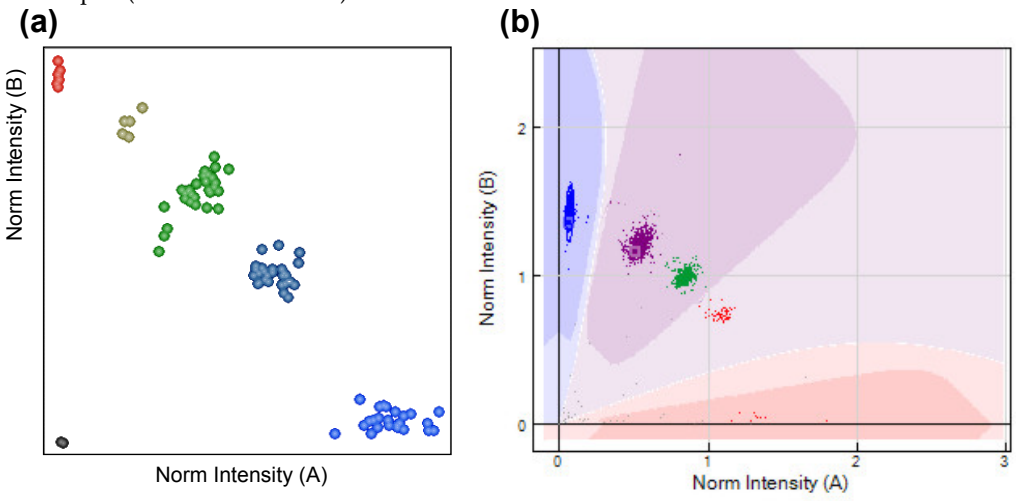

Figure 2. Examples of genotyping a binary SNP with five possible allelic states. (**a**) Genotyping using KASP (96 samples) and (**b**) an Infinium genotyping array (>2000 samples). Both genotyping assays produce two intensity signals, one for each allele, that are plotted on the X- and Y-axis. Samples with equal allel copy number are clustered in the five possible states (nulliplex to quadruplex).

A drawback for the efficient implementation of this kind of SNP genotyping assays is that the genotype quality scores of SNP assays are influenced by the number of SNPs present in the flanking probe sequence (MYLES *et al.* 2010). The highly polymorphic nature of the potato genome, with on average one SNP every ~20 bp, thus represents a challenge. The results of our GBS approach, with sequence variants discovered in a broad range of potato cultivars, calculated minor allele frequencies (MAF) and known flanking sequence context, however provides the required information for the design of both low- and high-density SNP genotyping assays.

A high-density potato Infinium SNP array (SolSTW-20k-Wageningen)

For more marker-dense applications a high-density Infinium SNP array (SolSTW-20K-Wageningen) has been developed (data not shown).By means of hierarchical cluster-analysis based on Kendall's tau rank correlation the complete set of >129,000 sequence variants discovered in this thesis was reduced to approximately 15,000 tag SNPs. Here, tag SNPs are defined as SNPs that effectively capture neighbouring variants trough LD, that is SNPs that effectively capture neighbouring variants trough LD. The redundancy of the complete set of sequence variants, with multiple variants tagging the same haplotypes, allowed us to choose tag SNPs with a minimal number of sequence variants flanking the target. Tag SNPs with low MAF or polymorphism information content (PIC) were not discarded but instead explicitly included, as we are interested in their frequencies and LD range in broader potato gene pools and in the contribution of rare haplotypes to phenotypic variation (SCHORK *et al.* 2009). Besides the mentioned ~15,000 tag SNPs, the SolSTW-20K-Wageningen array covers a number of chloroplast tag SNPs identified in our study, and ~4,500 SNPs identified by (HAMILTON *et al.* 2011) that overlap with the potato Illumina SolCAP SNP array (http://solcap.msu.edu/potato_infinium.shtml). Compared to the SNPs identified in this thesis, these SolCAP SNPs display a more dispersed coverage of the potato genome, but mainly represent common SNPs with a relatively high MAF.

A wide range of over 2,000 potato cultivars, accessions and breeding lines (both tetraploid and diploid) have been hybridized to the SolSTW-20K-Wageningen array. The resulting data are being analysed and will be used to answer a number of very interesting questions in population-genetics and plant breeding. It is anticipated that the SNP array will be efficient for applications like genome wide diversity- and association analysis and genomic selection (GS), and will elevate marker assisted breeding (MAB) in potato to a next level. Furthermore, the 84 cultivars of Chapter 4 in this thesis have been genotyped with the SNP array and these data can be used to directly compare a GBS method in potato to an array-based genotyping method.

High-throughput genotyping strategies, GBS or SNP arrays?

Compared to GBS, a major drawback of high-throughput SNP genotyping technologies is that only known SNPs are assayed. The assayed SNP markers are thus specific to the population for which they were developed, and genotyping of broader populations will be biased towards alleles present in the original survey. The sequence variants discovered in this thesis,

and the subset that can be assayed using the SolSTW-20K-Wageningen SNP array were identified mainly in commercial cultivars. For the analysis of wild or landrace potato cultivars an ascertainment bias has to be anticipated, as demonstrated for example in SNP assays of rice and barley (MORAGUES *et al.* 2010; THOMSON *et al.* 2012). In GBS on the other hand, genotyping of the same target regions in other populations can be achieved with further sequencing runs, accurately representing the alleles comprised by new populations and thus avoiding this bias towards the sequence variants in the originally surveyed population (DAVEY *et al.* 2011). With the rapid increasing throughput of sequencing, the automation of sample preparation and enrichment, and the efficient indexing of hundreds of individual cultivars, it is thus anticipated that there will be interest in moving rapidly towards GWAS and GS studies using direct GBS instead of genotyping arrays.

APPLICATIONS OF THE DISCOVERED DNA VARIANTS

A major limitation in genetic studies of potato has been the suboptimal number of available genetic markers, representative for the diverse potato gene pool. With the discovery of over 129,000 sequence variants in a world-wide set of cultivated potato germplasm, we hope to have alleviated this limitation. The discovered variants and genotyping results can be directly used for a diverse set of applications like diversity- and association analysis, haplotype-sharing and recombination analysis, genetic map development, and cultivar identification.

Low-density SNP assays

Besides the large scale SolSTW-20K-Wageningen fixed genotyping arrays, small panels of SNPs (or other sequence variants such as microsatellite markers) can be selected from the identified sequence variant repository, enabling both more traditional and high-throughput genotyping assays to be carried out on a small marker density scale, if genotyping of thousands of markers is not required by the application. We have for example developed KASP assays that have been used to validate the GBS results of Chapter 4, to validate marker-trait associations, to aid in the selection of near homozygous potato clones and for cultivar identification.

Cultivar identification

Cultivar identification is an important issue in potato. In the Potato pedigree database (VAN BERLOO *et al.* 2007), there are over 8,000 potato cultivar and progenitor lines. These vegetatively propagated genotypes are frequently confused due to the existence of multiple synonyms and homonyms and due to accidental substitution of material in for example propagation fields. In Chapter 3 we found that around 5% of the genotypes DNA we used might not represent the genuine cultivars. Currently 9 to 24 SSR loci are considered to be sufficient for genetic identification of most potato cultivars and landraces (GHISLAIN *et al.* 2009; REID *et al.* 2009). Standardization of allele sizes and estimation of allele copy number for SSRs is, although possible (ESSELINK *et al.* 2004), is still problematic and therefore alleles are commonly scored as either present or absent. Moreover, SSR genotyping is difficult to multiplex and SSR alleles are not always identical-by-descent. Given the low information content of SNPs, compared to SSRs a higher number of SNP markers are required to reach

similar resolution in genetic identification. The set of 270 KASP markers we used for validation of the GBS results of Chapter 4 can be used to determine allele copy numbers with high confidence. SNPs from this set have been selected from SNPs with a relatively high minor allele frequency (0.15 ≤MAF≤ 0.35) in the potato cultivar population and with a low mutual correlation, and are suitably distributed along the potato genome. These high-throughput SNP markers are thus an excellent tool for cultivar genotyping and can make the identification of potato cultivars more robust and convenient.

Marker-trait association

As an example for application of GBS for GWAS the identified sequence variants and genotype data were tested in a marker-trait association analysis with plant maturity and tuber flesh colour (Chapter 4). This led to the identification of alleles accounting for significant phenotypic variation in these traits and demonstrates that GWAS is feasible for common alleles of major effect at the current map resolution, and by using only a limited number of samples. In line with the results of Chapter 2, the association analysis effectively identified variation in and near the *CHY2* candidate gene as involved with the flesh colour QTL on chromosome 3. The *CHY2* gene has previously been identified as candidate gene for the well-known Y-locus involved in tuber flesh colour on chromosome 3 (BROWN *et al.* 2006). Also the QTL identified for early plant maturity in Chapter 4 is consistent with earlier reports of a major chromosome 5 QTL for this trait (BRADSHAW *et al.* 2004; VAN ECK and E. 1996; VISKER *et al.* 2003). In the analysis of chapter 4, the QTL region for the early plant maturity trait covered around 18 genes that could have causal influence on the trait. Recombinant analysis and complementation studies have been conducted and resulted in identification of the causal gene (BACHEM 2011), and some of the most significant markers identified in chapter 4 tag this gene.

The sequence variants discovered in Chapters 2 and 3 were genotyped in much larger potato germplasm populations, better suited to detect marker-trait associations for more complex traits. In these chapters we investigated the genetics of potato flesh colour and starch phosphate content. Extensive research has been performed on these and other potato quality traits and the candidate genes previously identified (BROWN *et al.* 2006; DIRETTO *et al.* 2006; LORBERTH *et al.* 1998; RITTE *et al.* 2006; RÖMER *et al.* 2002; THORUP *et al.* 2000; WERIJ *et al.* 2011). It is however the first time these candidate genes are analysed for the broad range of alleles present in the elite potato genepool. The association analysis led to the identification of previously unknown alleles related to the investigated traits, whose effects were further verified in di- and tetraploid mapping populations containing the relevant alleles.

Association analysis can be performed with either unphased sequence variants (Chapter 4) or haplotype markers (Chapters 2 and 3). Using haplotype-based association analysis in Chapters 2 and 3, with only a limited number of tagSNPs representing all haplotypes at the loci, we were able to assess the significance of all haplotype effects simultaneously in a multivariate model. In single-marker SNP association analysis such as performed in Chapter 4, there are often more (redundant) SNP markers then samples and association analysis is done repetitively on a SNP-by-SNP basis. The detection of alleles significantly associated with QTL

for haplotypes without a single unique tag SNP is difficult in such cases. The phenotypic effect of such a multi-marker defined haplotype will be underestimated by single-marker SNPs, and can result in an unexplained (missing) proportion of the heritability (BERGELSON and ROUX 2010). Although this inability to tag each haplotype by a single SNP can be solved using statistical models that test for interactions among single-marker SNPs, we suggest that identification of haplotypes can improve the identification of a causal haplotype with its phenotypic effect. Furthermore, haplotypes are essential to determine diversity parameters like allele richness and genotype composition, and to infer the biological context of alleles, because they help identify distinct variants (alleles) of genes.

We assume that identification of haplotypes and phasing of the sequence variants identified in Chapter 4 can be achieved. Multiple sequence variants present on the same sequencing read originate from the same allele and are thus in coupling phase. Given the high nucleotide diversity index in potato multiple phased variants per read are expected. With each read starting and ending at a unique position, and by using 'read-backed phasing' the local sequence variation can thus be interpreted into haplotypes within a small window of a continuously covered regions, where there is sequence variation but little or no confounding recombination. In combination with statistical LD phasing, larger haplotype blocks might even be identified.

Discovery of deleterious mutant alleles

The observed potato nucleotide diversity index in Chapter 4 translated into a density of ≈1 SNP/80 bp of noncoding sequence and ≈1 SNP/140 bp of coding sequence between two randomly selected homologous alleles. This allows predictions of the nature and distribution of sequence variants in all the estimated 39,000 potato genes, that is, in a study 50-fold larger. Based on the sample of 333 alleles of Chapter 4, we estimate that there are over 40 million sequence variants in the potato genome, with close to 1.5 million coding SNPs and 500 thousand non-synonymous SNPs. This is a tremendous resource for identifying functional allelic difference.

Although mutant alleles may accumulate in polyploid populations more quickly than in diploids (OTTO 2007) and potato is thought to have a high genetic load (VAN ECK *et al.* 1994), spontaneous knockout or reduction-of-function mutations are expected to be relatively rare. The alleles discovered in this thesis by screening the natural variation have been filtered by (natural) selection. In chapter 4, only around 4% of the sequenced genes showed allele with premature stop codon mutations. These include stop codons in candidate genes like α-Amylase (AMY3), the two known Solanidine galactosyl-transferase genes (SGT1 and SGT2) and Polyphenol oxidase (PPO) genes. Some of these mutations are quite rare, like the SGT2 loss-of-function mutation that occurs in only three cultivars, while others are more common. Ironically, the SGT2 loss-of-function mutation is found in three typical starch cultivars, where glycoalkaloid accumulation, for which SGT2 is a candidate gene, is considered of less importance.

DISCOVERY OF NOVEL DNA SEQUENCE VARIANTS BY EMS MUTAGENESIS

We used a mutagenesis approach to identify novel mutant alleles in genes involved in starch metabolism and carotenoid biosynthesis *(LCYe*, *ZEP*, *PWD*, *GWD*, *PAIN-1*, *SBEI*, *SBEII* and *SSSIII*) in Chapter 5. These candidate genes were screened in EMS-mutagenized populations using high resolution melting (HRM) analysis.

Evidence of causal effects of candidate genes usually requires mutants homozygous for the loss-of-function allele. To obtain these homozygous mutants with potential phenotypes we are selecting and crossing plants harbouring the discovered mutations. The first homozygous mutants, both of loss-of-function alleles and alleles predicted to reduce enzyme activity, have been identified and are being evaluated at the molecular and phenotypic level. For the identification of subtle quantitative effects of the alleles with novel amino-acids changes predicted to change enzyme activity, mutant alleles might need to be evaluated in more isogenic and diverse backgrounds to reduce genetic variation and to obtain more precise estimates of gene effects. Crossing the mutants to different genetic backgrounds will also be helpful to reduce the genetic damage caused by EMS mutagenesis at linked loci.

Advantage of screening a mutagenized M_1 generation

Individuals arising from mutagenized seeds (the M_1 generation) are chimeric. To avoid sampling chimeric tissue, commonly the M_2 generation of selfed M_1 plants is used for mutation screening. The requirement of this second non-chimeric M_2 generation for screening makes it time consuming. Furthermore, in diploid potato gametophytic incompatibility systems are active (EIJLANDER *et al.* 1997) that cause difficulties in the selfing of a mutagenized M_1 generation.

In traditional M_2 screening experiments, to avoid redundancy, it is only efficient to screen a limited number of plants per M_2 family and regularly only a single M_2 plant is analysed (BOTTICELLA *et al.* 2011; SLADE *et al.* 2005; TILL *et al.* 2007). Assuming normal segregation, this leads to a 25% chance that a specific mutation is lost in a M_2 plant produced by a selfed M_1 plant and a 50% chance that the mutation is lost if the M_1 plant has been crossed to a non-mutagenized donor plant. A lot of mutations in chimeric plants do however not segregate as expected. Depending on the meristem layer in which the original mutation is present, germline transmission and segregation in the M_2 generation can be distorted or absent. Furthermore, mutations genetically coupled to deleterious mutations will show distorted segregation in the M_2 generation. Mutations segregating at a (severely) reduced rate are therefore more likely to get lost if screening is commenced at the M_2 generation.

An advantage of screening the M_1 generation of a seed-treated EMS population is that no reproductive selection has taken place. Mutations detected in M_1 with distorted segregation in the M_2 generation can be followed and selected for in a large group of M_2 family plants. In Chapter 5 we show this can increase the number of potentially obtainable mutations. In the M_2 families of Chapter 5 we found mutations overrepresented more than three times, but also

mutations underrepresented more than 60 times, and only one out of four mutations segregated in the expected 1:1 ratio. Had we started screening at the M_2 generation and used only a limited number of M_2 family plants, two out of four mutations were likely to have gone unnoticed. However, we screened a large population of about 48 M_2 family plants per identified M_1 mutation and if necessary, this could have been extended to as many M_2 family plants as needed to identify a M_2 plant with the mutation.

Compared to M_2 screening, only a short developmental time is required for M_1 screening and the redundancy in screening M_2 families can be avoided. Having the facilities, many medium-sized M_1 plants can be grown simultaneously. Alternatively, high-throughput systems like the Ice-Cap method can be employed to grow plants on 96-well spin column plates and directly harvest the tissue for DNA isolation from these plates (CLARK and KRYSAN 2007; KRYSAN 2004). For vegetative propagation, *in vitro* explants can be made or in case of vegetatively propagated species like potato, tubers can be collected. For short-lived mutant populations, M_1 screening can only be performed for a relatively small number of genes since plants with mutations of interest have to be identified and propagated in a short time. Similar problems occur for example in M_2 population screening by iTILLING (BUSH and KRYSAN 2010), but a new EMS M_1 population is more quickly re-generated. For long-lived mutant populations like tree species or vegetatively propagated species like potato, whole-genome screening of seed-treated M_1 populations by MPS might even be applied.

Perspectives on genome-wide screening of a (M_1) population using MPS

Because of the reliability of EMS as a mutagen, the probability of success in recovering a loss-of-function mutation can be calculated in advance (MCCALLUM *et al.* 2000). In *Arabidopsis thaliana* 5% of the mutations induced by EMS in coding regions result in premature termination of the gene product (GREENE *et al.* 2003). The chance to obtain a stop codon mutation is variable over codon positions. A codon coding for tryptophan (TGG) for example has a higher chance to mutate into a stop codon (TGA or TAG) than other amino acid codons . We took this into account during the amplicon design for mutation screening in Chapter 5. Presumably because of this, we found a somewhat higher percentage of stop codon mutations in the coding sequence (~8%) than found in Arabidopsis. Assuming however that also in potato 5% of all coding mutations are stop codon mutations, and with a total coding sequence length of 36 Mb dispersed across 39,000 potato genes (XU *et al.* 2011), each M_1 plant of Chapter 5 is expected to contain about 554 coding sequence mutations (1 mutation/65 kb) and 28 loss-of-function mutations. This suggests that for all potato genes a knockout mutation can be identified by using a population of approximately 1400 M_1 plants. Besides the knockouts, this collection would contain a whole range of approximately 20 coding sequence mutations per gene. When plants are beforehand selected for M_1-fertility, a large number of these mutations are expected to be "accessible" mutations that can be bred into genetically stable M_2 plants for gene function analysis. Current MPS capacity of the Illumina HiSeq is approximately 300 Gb per flow cell. In a diploid, a read depth of 10-15× implies a 95-99% chance to sequence a mutant allele two or three times. Sequencing the complete 36 Mb coding sequence at this depth for 1400 plants (≈ 500-750 Gb) is thus reachable and could in the future make

mutagenesis screening a largely *in silico* procedure, with searchable online databases from which mutants of interest can be ordered, similar to for example Arabidopsis (ALONSO *et al.* 2003).

CONCLUDING REMARKS

In this thesis two parallel routes have been followed with the aim to sample and to create genetic diversity in potato germplasm, on the one hand by the screening of induced mutant populations by HRM, and on the other hand by screening of natural sequence variants by sequencing. A first conclusion is that both methods have resulted in the successful identification of novel alleles. HRM mutant screening resulted in the identification of five mutant plants with stop codons in candidate genes important to potato breeding. Sequencing of existing germplasm allowed the identification of functionally different *CHY2*, *ZEP* and *StGWD* alleles as well as the discovery of premature stop codons in 43 potato genes.

A second conclusion is that within tetraploid potato germplasm the number of recessive loss-of-function mutations is not as high as anticipated in the project proposal. For many genes we have not discovered loss-of-function mutations, and therefore chemical mutagenesis will remain an important method to acquire this type of mutants.

Main accomplishments of this thesis

In a broader context, this thesis has made key contributions to potato breeding and, more generally, the field of theoretical and applied genetics as follows:

- The discovery of *StGWD* alleles with a quantitative effect on starch phosphate content in commercial potato cultivars and the identification of a *StZEP* allele causing orange tuber flesh colour.
- The application of MPS to indexed and sequence-enriched libraries of 83 genetically diverse tetraploid potato cultivars and a single monoploid accession, leading to the identification and genotyping of over 129,000 sequence variants in 2.1 Mb of the potato genome and covering >800 target genes.
- Proof of the GBS concept for autotetraploid species using first- and next-generation sequencing and establishment of an accuracy threshold on MPS sequence depth ensuring reliable quantitative genotype calls.
- Development of both small- (KASP) and large-scale (Infinium) SNP assays for genotyping potato.
- Establishment of EMS-induced mutation screening in M_1 generations as an effective alternative to M_2 mutation screening for species for which selfing of the M_1 generation is difficult.
- Discovery of loss-of-function alleles in candidate genes of existing potato germplasm.
- Development of an EMS mutagenized population and identification of allelic series, including loss-of-function alleles, for six candidate genes known to influence potato quality, and the genetic stabilisation of these mutants.

References

ADZHUBEI, I. A., S. SCHMIDT, L. PESHKIN, V. E. RAMENSKY, A. GERASIMOVA *et al.*, 2010 A method and server for predicting damaging missense mutations. Nature Methods **7:** 248-249.

ALLENDER, C. J., and G. J. KING, 2010 Origins of the amphiploid species Brassica napus L. investigated by chloroplast and nuclear molecular markers. BMC Plant Biology **10**.

ALONSO, J. M., and J. R. ECKER, 2006 Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. **7:** 524-536.

ALONSO, J. M., A. N. STEPANOVA, T. J. LEISSE, C. J. KIM, H. CHEN *et al.*, 2003 Genome-Wide Insertional Mutagenesis of Arabidopsis thaliana. Science **301:** 653-657.

AMES, M., and D. M. SPOONER, 2008 DNA from herbarium specimens settles a controversy about origins of the European potato. American Journal of Botany **95:** 252-257.

ANDERSEN, J. R., and T. LÜBBERSTEDT, 2003 Functional markers in plants. Trends in Plant Science **8:** 554-560.

ANDRE, C. M., M. OUFIR, C. GUIGNARD, L. HOFFMANN, J. F. HAUSMAN *et al.*, 2007 Antioxidant profiling of native Andean potato tubers (Solanum tuberosum L.) reveals cultivars with high levels of β-carotene, α-tocopherol, chlorogenic acid, and petanin. Journal of Agricultural and Food Chemistry **55:** 10839-10849.

ANITHAKUMARI, A. M., J. TANG, H. J. VAN ECK, R. G. VISSER, J. A. LEUNISSEN *et al.*, 2010 A pipeline for high throughput detection and mapping of SNPs from EST databases. Molecular breeding : new strategies in plant improvement **26:** 65-75.

ATWELL, S., Y. S. HUANG, B. J. VILHJALMSSON, G. WILLEMS, M. HORTON *et al.*, 2010 Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature **465:** 627-631.

BACHEM, C. W. B., R. J. F. J. OOMEN and R. G. F. VISSER, 1998 Transcript Imaging with cDNA-AFLP: A Step-by-Step Protocol. Plant Molecular Biology Reporter **16:** 157-173.

BACHEM, C. W., 2011 Cloning of the gene responsible for the earliness QTL in potato: Regulation of tuberisation and plant life-cycle is linked and controlled by genes associated to the circadian clock. Abstract, 8th Solanaceae and 2nd Cucurbitaceae Joint Conference, SOL, Kobe, Japan.

BAIRD, N. A., P. D. ETTER, T. S. ATWOOD, M. C. CURREY, A. L. SHIVER *et al.*, 2008 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PloS one **3:** e3376.

BALLVORA, A., M. R. ERCOLANO, J. WEIß, K. MEKSEM, C. A. BORMANN *et al.*, 2002 The R1 gene for potato resistance to late blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. The Plant Journal **30:** 361-371.

BERGELSON, J., and F. ROUX, 2010 Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. **11:** 867-879.

BERTONE, P., V. TRIFONOV, J. S. ROZOWSKY, F. SCHUBERT, O. EMANUELSSON *et al.*, 2006 Design optimization methods for genomic DNA tiling arrays. Genome Research **16:** 271-281.

BINDING, H., R. NEHLS, O. SCHIEDER, S. K. SOPORY and G. WENZEL, 1978 Regeneration of mesophyll protoplasts isolated from dihaploid clones of Solanum tuberosum. Physiol. Plant. **43:** 52-54.

BINO, R. J., C. H. R. DE VOS, M. LIEBERMAN, R. D. HALL, A. BOVY *et al.*, 2005 The light-hyperresponsive high pigment-2dg mutation of tomato: Alterations in the fruit metabolome. New Phytologist **166:** 427-438.

BONIERBALE, M. W., R. L. PLAISTED and S. D. TANKSLEY, 1988 RFLP Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato. Genetics **120:** 1095-1103.

BOTTICELLA, E., F. SESTILI, A. HERNANDEZ-LOPEZ, A. PHILLIPS and D. LAFIANDRA, 2011 High resolution melting analysis for the detection of EMS induced mutations in wheat SbeIIa genes. BMC Plant Biology **11**.

BRADSHAW, J. E., B. PANDE, G. J. BRYAN, C. A. HACKETT, K. MCLEAN *et al.*, 2004 Interval mapping of quantitative trait loci for resistance to late blight [Phytophthora infestans (Mont.) de Bary], height and maturity in a tetraploid population of potato (Solanum tuberosum subsp tuberosum). Genetics **168:** 983-995.

BRADSHAW, J. E., and G. RAMSAY, 2005 Utilisation of the Commonwealth Potato Collection in potato breeding. Euphytica **146:** 9-19.

BREITBAUPT, D. E., and A. BAMEDI, 2002 Carotenoids and carotenoid esters in potatoes (Solanum tuberosum L.): New insights into an ancient vegetable. Journal of Agricultural and Food Chemistry **50:** 7175-7181.

BROWN, A. H. D., 1989 Core collections: a practical approach to genetic resources management. Genome **31:** 818-824.

BROWN, C. R., C. G. EDWARDS, C. P. YANG and B. B. DEAN, 1993 Orange flesh trait in potato: Inheritance and carotenoid content. J. Amer. Soc. Hort. Sci. **118:** 145-150.

BROWN, C. R., T. S. KIM, Z. GANGA, K. HAYNES, D. DE JONG *et al.*, 2006 Segregation of total carotenoid in high level potato germplasm and its relationship to beta-carotene hydroxylase polymorphism. American Journal of Potato Research **83:** 365-372.

BROWN, C. R., D. CULLEY, M. BONIERBALE and W. AMORÓS, 2007 Anthocyanin, carotenoid content, and antioxidant values in native South American potato cultivars. Hortscience **42:** 1733-1736.

BROWN, C. R., 2008 Breeding for phytonutrient enhancement of potato. American Journal of Potato Research **85:** 298-307.

BUNTJER, J. B., A. P. SØRENSEN and J. D. PELEMAN, 2005 Haplotype diversity: the link between statistical and biological association. Trends in Plant Science **10:** 466-471.

BURGOS, G., E. SALAS, W. AMOROS, M. AUQUI, L. MUÑOA *et al.*, 2009 Total and individual carotenoid profiles in Solanum phureja of cultivated potatoes: I. Concentrations and relationships as determined by spectrophotometry and HPLC. Journal of Food Composition and Analysis **22:** 503-508.

BUSH, S. M., and P. J. KRYSAN, 2010 iTILLING: A Personalized Approach to the Identification of Induced Mutations in Arabidopsis. Plant physiology **154:** 25-35.

CALDWELL, D. G., N. MCCALLUM, P. SHAW, G. J. MUEHLBAUER, D. F. MARSHALL *et al.*, 2004 A structured mutant population for forward and reverse genetics in Barley (Hordeum vulgare L.). Plant Journal **40:** 143-150.

CASTILLO, A., G. DORADO, C. FEUILLET, P. SOURDILLE and P. HERNANDEZ, 2010 Genetic structure and ecogeographical adaptation in wild barley (Hordeum chilense Roemer et Schultes) as revealed by microsatellite markers. BMC Plant Biology **10**.

CELIS GAMBOA, B. C., 2002 The life cycle of the potato (Solanum tuberosum L.): From physiology to genetics. 186.

CHARLESWORTLI, B., and D. CHARLESWORTH, 1998 Some evolutionary consequences of deleterious mutations. Genetica **102-103:** 3-19.

CHAWADE, A., P. SIKORA, M. BRAUTIGAM, M. LARSSON, V. VIVEKANAND *et al.*, 2010 Development and characterization of an oat TILLING-population and identification of mutations in lignin and beta-glucan biosynthesis genes. BMC Plant Biology **10:** 86.

CHING, A., K. S. CALDWELL, M. JUNG, M. DOLAN, O. S. SMITH *et al.*, 2002 SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. Bmc Genetics **3**.

CINGOLANI, P., A. PLATTS, L. WANG, M. COON, T. NGUYEN *et al.*, 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**.

CLARK, A. G., 2004 The role of haplotypes in candidate gene studies. Genetic Epidemiology **27:** 321-333.

CLARK, K., and P. KRYSAN, 2007 Protocol: An improved high-throughput method for generating tissue samples in 96-well format for plant genotyping (Ice-Cap 2.0). Plant methods **3:** 8.

CLOSE, T., P. BHAT, S. LONARDI, Y. WU, N. ROSTOKS *et al.*, 2009 Development and implementation of high-throughput SNP genotyping in barley. BMC genomics **10:** 582.

COMAI, L., and S. HENIKOFF, 2006 TILLING: Practical single-nucleotide mutation discovery. Plant Journal **45:** 684-694.

CRAIG, D. W., J. V. PEARSON, S. SZELINGER, A. SEKAR, M. REDMAN *et al.*, 2008 Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods **5:** 887-893.

CRAIG, S. A. S., C. C. MANINGAT, P. A. SEIB and R. C. HOSENEY, 1989 Starch paste clarity. Cereal chemistry. **66:** 173-182.

CRONN, R., A. LISTON, M. PARKS, D. S. GERNANDT, R. SHEN *et al.*, 2008 Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research **36:** e122.

D'HOOP, B. B., M. J. PAULO, K. KOWITWANICH, M. SENGERS, R. G. F. VISSER *et al.*, 2010 Population structure and linkage disequilibrium unravelled in tetraploid potato. Theoretical and Applied Genetics **121:** 1151-1170.

D'HOOP, B. B., M. J. PAULO, R. A. MANK, H. J. VAN ECK and F. A. VAN EEUWIJK, 2008 Association mapping of quality traits in potato (Solanum tuberosum L.). Euphytica **161:** 47-60.

DANIELL, H., 2004 Chloroplast genetic engineering for improved agronomic traits and molecular farming, using various selection systems. In Vitro Cellular & Developmental Biology-Animal **40:** 18a-18a.

DAVEY, J. W., P. A. HOHENLOHE, P. D. ETTER, J. Q. BOONE, J. M. CATCHEN *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **12:** 499-510.

DE BAKKER, P. I. W., R. YELENSKY, I. PE'ER, S. B. GABRIEL, M. J. DALY *et al.*, 2005 Efficiency and power in genetic association studies. Nature Genetics **37:** 1217-1223.

DE JONG, W., A. FORSYTH, D. LEISTER, C. GEBHARDT and D. C. BAULCOMBE, 1997 A potato hypersensitive resistance gene against potato virus X maps to a resistance gene cluster on chromosome 5. Theoretical and Applied Genetics **95:** 246-252.

DE KOEYER, D., K. DOUGLASS, A. MURPHY, S. WHITNEY, L. NOLAN *et al.*, 2009 Application of high-resolution DNA melting for genotyping and variant scanning of diploid and autotetraploid potato. Molecular Breeding **25:** 67-90.

DE NETTANCOURT, D., and M. DIJSTRA, 1969 Starch accumulation in the microspores of a solanum species and possible implications in mutation breeding. American Journal of Potato Research **46:** 239-242.

DE NOOY, W., A. MRVAR and V. BATAGELJ, 2005 Exploratory Social Network Analysis with Pajek.

DE VRIES-UIJTEWAAL, E., L. J. W. GILISSEN, E. FLIPSE, K. SREE RAMULU, W. J. STIEKEMA *et al.*, 1989 Fate of introduced genetic markers in transformed root clones and regenerated plants of monohaploid and diploid potato genotypes. Theoretical and Applied Genetics **78:** 185-193.

DE VRIES, S. E., M. A. FERWERDA, A. E. H. M. LOONEN, L. P. PIJNACKER and W. J. FEENSTRA, 1987 Chromosomes in somatic hybrids between Nicotiana plumbaginifolia and a monoploid potato. Theoretical and Applied Genetics **75:** 170-176.

DIRETTO, G., R. TAVAZZA, R. WELSCH, D. PIZZICHINI, F. MOURGUES *et al.*, 2006 Metabolic engineering of potato tuber carotenoids through tuber-specific silencing of lycopene epsilon cyclase. BMC Plant Biology **6**.

DUCKHAM, S. C., R. S. T. LINFORTH and I. B. TAYLOR, 1991 Abscisic acid-deficient mutants at the aba gene locus of Arabidopsis thaliana are impaired in the epoxidation of zeaxanthin. Plant Cell Environ. **14:** 601-606.

DURBIN, R., and H. LI, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:** 1754-1760.

DURBIN, R., H. LI, B. HANDSAKER, A. WYSOKER, T. FENNELL *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:** 2078-2079.

EIJLANDER, R., M. S. RAMANNA and E. JACOBSEN, 1997 Selection of vigorous and fertile S-homo- and heterozygous tester clones from self-incompatible diploid potato, Solanum tuberosum L. Euphytica **97:** 97-111.

ELLENBY, C., 1952 Resistance to the potato root eelworm, Heterodera rostochiensis Wollenweber [2]. Nature **170:** 1016.

ELSHIRE, R. J., J. C. GLAUBITZ, Q. SUN, J. A. POLAND, K. KAWAMOTO *et al.*, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PloS one **6:** e19379.

ESSELINK, G. D., H. NYBOM and B. VOSMAN, 2004 Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting—peak ratios) method. TAG Theoretical and Applied Genetics **109:** 402-408.

FRANCIA, E., G. TACCONI, C. CROSATTI, D. BARABASCHI, D. BULGARELLI *et al.*, 2005 Marker assisted selection in crop plants. Plant Cell, Tissue and Organ Culture **82:** 317-342.

FRUWIRTH, C., 1912 Zur Züchtung der Kartoffel. Deutsche Landwirtschaftliche Presse **39:** 565-567.

GABRIEL, S. B., S. F. SCHAFFNER, H. NGUYEN, J. M. MOORE, J. ROY *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225-2229.

GADY, A. L. F., F. W. K. HERMANS, M. H. B. J. VAN DE WAL, E. N. VAN LOO, R. G. F. VISSER *et al.*, 2009 Implementation of two high through-put techniques in a novel application: Detecting point mutations in large EMS mutated plant populations. Plant methods **5**.

GALPAZ, N., Q. WANG, N. MENDA, D. ZAMIR and J. HIRSCHBERG, 2008 Abscisic acid deficiency in the tomato mutant high-pigment 3 leading to increased plastid number and higher fruit lycopene content. Plant Journal **53:** 717-730.

GANAL, M. W., G. DURSTEWITZ, A. POLLEY, A. BÉRARD, E. S. BUCKLER *et al.*, 2011 A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. PloS one **6:** e28334.

GAZZANI, S., A. R. GENDALL, C. LISTER and C. DEAN, 2003 Analysis of the molecular basis of flowering time variation in Arabidopsis accessions. Plant physiology **132:** 1107-1114.

GEBHARDT, C., X. CHEN and F. SALAMINI, 2001 A potato molecular-function map for carbohydrate metabolism and transport. Theoretical and Applied Genetics **102:** 284-295.

GHISLAIN, M., J. NÚÑEZ, M. DEL ROSARIO HERRERA, J. PIGNATARO, F. GUZMAN *et al.*, 2009 Robust and highly informative microsatellite-based genetic identity kit for potato. Molecular Breeding **23:** 377-388.

GNIRKE, A., A. MELNIKOV, J. MAGUIRE, P. ROGOV, E. M. LEPROUST *et al.*, 2009 Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology **27:** 182-189.

GONZÁLEZ-SCHAIN, N. D., M. DÍAZ-MENDOZA, M. ŻURCZAK and P. SUÁREZ-LÓPEZ, 2012 Potato CONSTANS is involved in photoperiodic tuberization in a graft-transmissible manner. The Plant Journal **70:** 678-690.

GOO, Y. M., T. W. KIM, S. H. HA, K. W. BACK, J. M. BAE *et al.*, 2009 Expression profiles of genes involved in the carotenoid biosynthetic pathway in yellow-fleshed potato cultivars (Solanum tuberosum L.) from South Korea. Journal of Plant Biology **52:** 49-55.

GORE, M., P. BRADBURY, R. HOGERS, M. KIRST, E. VERSTEGE *et al.*, 2007 Evaluation of Target Preparation Methods for Single-Feature Polymorphism Detection in Large Complex Plant Genomes. Crop Sci. **47:** S-135-S-148.

GORE, M. A., M. H. WRIGHT, E. S. ERSOZ, P. BOUFFARD, E. S. SZEKERES *et al.*, 2009 Large-Scale Discovery of Gene-Enriched SNPs. Plant Gen. **2:** 121-133.

GORT, G., and F. VAN EEUWIJK, 2012 Review and simulation of homoplasy and collision in AFLP. Euphytica **183:** 389-400.

GRATTAPAGLIA, D., O. JUNIOR, M. KIRST, B. DE LIMA, D. FARIA *et al.*, 2011 High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC Plant Biology **11:** 65.

GREENE, E. A., C. A. CODOMO, N. E. TAYLOR, J. G. HENIKOFF, B. J. TILL *et al.*, 2003 Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in Arabidopsis. Genetics **164:** 731-740.

GRIFFIN, P., C. ROBIN and A. HOFFMANN, 2011 A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. BMC Biology **9:** 19.

HAASE, N. U., and J. PLATE, 1996 Properties of potato starch in relation to varieties and environmental factors. Starch/Staerke **48:** 167-171.

HAMILTON, J. P., C. N. HANSEY, B. R. WHITTY, K. STOFFEL, A. N. MASSA *et al.*, 2011 Single nucleotide polymorphism discovery in elite north american potato germplasm. BMC genomics **12**.

HANDELMAN, G. J., E. A. DRATZ, C. C. REAY and F. J. G. M. VAN KUIJK, 1988 Carotenoids in the human macula and whole retina. Investigative Ophthalmology and Visual Science **29:** 850-855.

HANNEMAN JR, R. E., and S. J. PELOQUIN, 1967 Croosability of 24-chromosome potato hybrids with 48-chromosome cultivars. European Potato Journal **10:** 62-73.

HANSON, M. R., 1989 Tracking down plant genes: Paths patterns, and footprints. Plant Cell **1:** 169-172.

HARISMENDY, O., R. TEWHEY, M. NAKANO, X. Y. WANG, C. PABON-PENA *et al.*, 2009 Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biology **10**.

HERMSEN, J. G. T., 1978 Genetics of self-compatibility in dihaploids of *Solanum tuberosum* L. 4. Linkage between an *S*-bearing translocation and a locus for virescens. Euphytica **27:** 381-384.

HILDERING, G. J., and J. H. VAN DER VEEN, 1966 The mutual independence of M1-fertility and mutant yield in EMS treated tomatoes. Euphytica **15:** 412-424.

HODGES, E., M. ROOKS, Z. XUAN, A. BHATTACHARJEE, D. BENJAMIN GORDON *et al.*, 2009 Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. **4:** 960-974.

HOOGKAMP, T. J. H., R. G. T. VAN DEN ENDE, E. JACOBSEN and R. G. F. VISSER, 2000 Development of amylose-free (amf) monoploid potatoes as new basic material for mutation breeding in vitro. Potato Research **43:** 179-189.

HOSAKA, K., 2004 Evolutionary pathway of T-type Chloroplast DNA in potato. American Journal of Potato Research **81:** 153-158.

HOSAKA, K., G. A. DE ZOETEN and R. E. HANNEMAN JR, 1988 Cultivated potato chloroplast DNA differs from the wild type by one deletion - evidence and implications. Theoretical and Applied Genetics **75:** 741-745.

HOSAKA, K., and R. E. HANNEMAN, 1998 Genetics of self-compatibility in a self-incompatible wild diploid potato species Solanum chacoense. 2. Localization of an S locus inhibitor (Sli) gene on the potato genome using DNA markers. Euphytica **103:** 265-271.

HOVENKAMP-HERMELINK, J. H. M., E. JACOBSEN, L. P. PIJNACKER, J. N. DE VRIES, B. WITHOLT *et al.*, 1988 Cytological studies on adventitious shoots and minitubers of a monoploid potato clone. Euphytica **39:** 213-219.

HOVENKAMP-HERMELINK, J. H. M., E. JACOBSEN, A. S. PONSTEIN, R. G. F. VISSER, G. H. VOS-SCHEPERKEUTER *et al.*, 1987 Isolation of an amylose-free starch mutant of the potato (*Solanum tuberosum* L.). TAG Theoretical and Applied Genetics **75:** 217-221.

HUANG, S., E. A. G. VAN DER VOSSEN, H. KUANG, V. G. A. A. VLEESHOUWERS, N. ZHANG *et al.*, 2005 Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato. The Plant Journal **42:** 251-261.

HUANG, X., X. WEI, T. SANG, Q. ZHAO, Q. FENG *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. **42:** 961-967.

HUTTEN, R. C. B., M. G. M. SCHIPPERS, J. G. T. HERMSEN and M. S. RAMANNA, 1994 Comparative performance of FDR and SDR progenies from reciprocal 4x-2x crosses in potato. TAG Theoretical and Applied Genetics **89:** 545-550.

IAEA, 1977 Manual on Mutation Breeding, 2nd edition. Technical Report Series **119**.

ISHIKAWA, T., Y. KAMEI, S. OTOZAI, J. KIM, A. SATO *et al.*, 2010 High-resolution melting curve analysis for rapid detection of mutations in a Medaka TILLING library. BMC Molecular Biology **11**.

IWANZIK, W., M. TEVINI, R. STUTE and R. HILBERT, 1983 Carotinoidgehalt und-zusammensetzung verschiedener deutscher Kartoffelsorten und deren Bedeutung für die Fleischfarbe der Knolle. Potato Research **26:** 149-162.

JACOBS, J. M. E., H. J. VAN ECK, P. ARENS, B. VERKERK-BAKKER, B. TE LINTEL HEKKERT *et al.*, 1995 A genetic map of potato (Solanum tuberosum) integrating molecular markers, including transposons, and classical markers. Theoretical and Applied Genetics **91:** 289-300.

JACOBSEN, E., 1980 Increase of diplandroid formation and seed set in 4x × 2x crosses in potatoes by genetical manipulation of dihaploids and some theoretical consequences. Z Pflanzenzu?chtg **85:** 110-121.

JACOBSEN, E., J. H. M. HOVENKAMP-HERMELINK, H. T. KRIJGSHELD, H. NIJDAM, L. P. PIJNACKER *et al.*, 1989 Phenotypic and genotypic characterization of an amylose-free starch mutant of the potato. Euphytica **44:** 43-48.

JANNINK, J.-L., A. J. LORENZ and H. IWATA, 2010 Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics **9:** 166-177.

JOBLING, S., 2004 Improving starch for food and industrial applications. Current Opinion in Plant Biology **7:** 210-218.

JOHNSON, G. C. L., L. ESPOSITO, B. J. BARRATT, A. N. SMITH, J. HEWARD *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nature Genetics **29:** 233-237.

KAPITONOV, V. V., S. TEMPEL and J. JURKA, 2009 Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene **448:** 207-213.

KENNY, E. M., P. CORMICAN, W. P. GILKS, A. S. GATES, C. T. O'DUSHLAINE *et al.*, 2011 Multiplex Target Enrichment Using DNA Indexing for Ultra-High Throughput SNP Detection. DNA Research **18:** 31-38.

KISS, M. M., L. ORTOLEVA-DONNELLY, N. REGINALD BEER, J. WARNER, C. G. BAILEY *et al.*, 2008 High-throughput quantitative polymerase chain reaction in picoliter droplets. Analytical Chemistry **80:** 8975-8981.

KLOOSTERMAN, B., D. DE KOEYER, R. GRIFFITHS, B. FLINN, B. STEUERNAGEL *et al.*, 2008 Genes driving potato tuber initiation and growth: Identification based on transcriptional changes using the POCI array. Functional and Integrative Genomics **8:** 329-340.

KLOOSTERMAN, B., M. OORTWIJN, J. UITDEWILLIGEN, T. AMERICA, R. DE VOS *et al.*, 2010 From QTL to candidate gene: Genetical genomics of simple and complex traits in potato using a pooling strategy. BMC genomics **11**.

KOBAYASHI, A., A. OHARA-TAKADA, S. TSUDA, C. MATSUURA-ENDO, N. TAKADA *et al.*, 2008 Breeding of potato variety "Inca-no-hitomi" with a very high carotenoid content. Breeding Science **58:** 77-82.

KORT, J., C. P. JASPERS and D. L. DIJKSTRA, 1972 Testing for resistance to pathotype C of Heterodera rostochiensis and the practical application of Solanum vernei-hybrids in the Netherlands. Ann Appl Biol **71:** 289-294.

KÖTTING, O., J. KOSSMANN, S. C. ZEEMAN and J. R. LLOYD, 2010 Regulation of starch metabolism: The age of enlightenment? Current Opinion in Plant Biology **13:** 321-329.

KOVALCHUK, I., O. KOVALCHUK and B. HOHN, 2000 Genome-wide variation of the somatic mutation frequency in transgenic plants. **19:** 4431-4438.

KRAAK, A., 1992 Industrial applications of potato starch products. Industrial Crops and Products **1:** 107-112.

KRINSKY, N. I., S. T. MAYNE and H. SIES, 2004 Carotenoids in Health and Disease. CRC Press, New York

KRUMM, N., P. H. SUDMANT, A. KO, B. J. O'ROAK, M. MALIG *et al.*, 2012 Copy number variation detection and genotyping from exome sequence data. Genome Research.

KRYSAN, P., 2004 Ice-Cap. A High-Throughput Method for Capturing Plant Tissue Samples for Genotype Analysis. Plant physiology **135:** 1162-1169.

KUROWSKA, M., A. DASZKOWSKA-GOLEC, D. GRUSZKA, M. MARZEC, M. SZURMAN *et al.*, 2011 TILLING - a shortcut in functional genomics. Journal of applied genetics **52:** 371-390.

LAM, H. Y. K., X. J. MU, A. M. STUTZ, A. TANZER, P. D. CAYTING *et al.*, 2010 Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. **28:** 47-55.

LÊ, S., J. JOSSE and F. HUSSON, 2008 FactoMineR: An R package for multivariate analysis. Journal of Statistical Software **25:** 1-18.

LI, L., J. STRAHWALD, H. R. HOFFERBERT, J. LÜBECK, E. TACKE *et al.*, 2005 DNA variation at the invertase locus invGE/GF is associated with tuber quality traits in populations of potato breeding clones. Genetics **170:** 813-821.

LIGHTBOURN, G., and R. VEILLEUX, 2003 Retrotransposon based markers to characterize somatic hybrids and assess variation induced by protoplast fusion of monoploid potato. Potatoes - Healthy Food for Humanity: International Developments in Breeding, Production, Protection and Utilization**:** 35-43.

LIJAVETZKY, D., J. CABEZAS, A. IBANEZ, V. RODRIGUEZ and J. MARTINEZ-ZAPATER, 2007 High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC genomics **8:** 424.

LOPEZ, A. B., J. VAN ECK, B. J. CONLIN, D. J. PAOLILLO, J. O'NEILL *et al.*, 2008 Effect of the cauliflower or transgene on carotenoid accumulation and chromoplast formation in transgenic potato tubers. Journal of Experimental Botany **59:** 213-223.

LORBERTH, R., G. RITTE, L. WILLMITZER and J. KOSSMANN, 1998 Inhibition of a starch-granule-bound protein leads to modified starch and repression of cold sweetening. Nature Biotechnology **16:** 473-477.

MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER *et al.*, 2010 Target-enrichment strategies for next-generation sequencing. **7:** 111-118.

MARIN, E., L. NUSSAUME, A. QUESADA, M. GONNEAU, B. SOTTA *et al.*, 1996 Molecular identification of zeaxanthin epoxidase of Nicotiana plumbaginifolia, a gene involved in abscisic acid biosynthesis and corresponding to the ABA locus of Arabidopsis thaliana. EMBO Journal **15:** 2331-2342.

MARTIN, C., R. NIGGEWEG and A. J. MICHAEL, 2004 Engineering plants with increased levels of the antioxidant chlorogenic acid. Nature Biotechnology **22:** 746-754.

MAY, B. P., and R. A. MARTIENSSEN, 2003 Transposon mutagenesis in the study of plant development. Critical Reviews in Plant Sciences **22:** 1-35.

MCCALLUM, C. M., L. COMAI, E. A. GREENE and S. HENIKOFF, 2000 Targeting Induced LocalLesions IN Genomes (TILLING) for Plant Functional Genomics. Plant physiology **123:** 439-442.

MCCOUCH, S. R., K. ZHAO, M. WRIGHT, C.-W. TUNG, K. EBANA *et al.*, 2010 Development of genome-wide SNP assays for rice. Breeding Science **60:** 524-535.

MCCUE, K. F., P. V. ALLEN, L. V. T. SHEPHERD, A. BLAKE, D. R. ROCKHOLD *et al.*, 2007 Manipulation and compensation of steroidal glycoalkaloid biosynthesis in potatoes. Proceedings of the VIth International Solanaceae Conference, Solanaceae VI: Genomics Meets Biodiversity**:** 343-349.

MCGRATH, J., R. SHAW, B. DE LOS REYES and J. WEILAND, 2004 Construction of a sugar beet BAC library from a hybrid with diverse traits. Plant Molecular Biology Reporter **22:** 23-28.

MCKENNA, A., M. HANNA, E. BANKS, A. SIVACHENKO, K. CIBULSKIS *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20:** 1297-1303.

MCMANUS, L. J., J. SASSE, C. K. BLOMSTEDT and G. BOSSINGER, 2006 Pollen treatment for mutation induction in Eucalyptus globulus ssp. globulus (Myrtaceae). Australian Journal of Botany **54:** 65-71.

MCMANUS, L. J., J. SASSE, C. K. BLOMSTEDT and G. BOSSINGER, 2007 The effects of ethyl methanesulfonate treatment on Eucalyptus pollen behaviour in vitro. Trees - Structure and Function **21:** 379-383.

MENDA, N., R. M. BUELS, I. TECLE and L. A. MUELLER, 2008 A community-based annotation framework for linking solanaceae genomes with phenomes. Plant physiology **147:** 1788-1799.

METZKER, M. L., 2005 Emerging technologies in DNA sequencing. Genome Research **15:** 1767-1776.

MEYER, S., A. NAGEL and C. GEBHARDT, 2005 PoMaMo--a comprehensive database for potato genome data. Nucleic Acids Research **33:** D666-670.

MICHAELS, S. D., Y. HE, K. C. SCORTECCI and R. M. AMASINO, 2003 Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **100:** 10102-10107.

MINOIA, S., A. PETROZZA, O. D'ONOFRIO, F. PIRON, G. MOSCA *et al.*, 2010 A new mutant genetic resource for tomato crop improvement by TILLING technology. BMC Research Notes **3**.

MOELLER, S. M., N. PAREKH, L. TINKER, C. RITENBAUGH, B. BLODI *et al.*, 2006 Associations between intermediate age-related macular degeneration and lutein and zeaxanthin in the Carotenoids in Age-Related Eye Disease Study (CAREDS): Ancillary study of the Women's Health Initiative. Archives of Ophthalmology **124:** 1151-1162.

MOH, C.-C., 1950 An analysis of seedling mutants (spontaneous, atomic bomb-radiation-, and X-Ray-induced) in barley and Durum wheat, pp.

MOOSE, S. P., and R. H. MUMM, 2008 Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement. Plant physiology **147:** 969-977.

MORAGUES, M., J. COMADRAN, R. WAUGH, I. MILNE, A. FLAVELL *et al.*, 2010 Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. TAG Theoretical and Applied Genetics **120:** 1525-1534.

MORRIS, W. L., L. DUCREUX, D. W. GRIFFITHS, D. STEWART, H. V. DAVIES *et al.*, 2004 Carotenogenesis during tuber development and storage in potato. Journal of Experimental Botany **55:** 975-982.

MULLARKEY, M., and P. JONES, 2000 Isolation and analysis of thermotolerant mutants of wheat. Journal of Experimental Botany **51:** 139-146.

MUTH, J., S. HARTJE, R. M. TWYMAN, H. R. HOFFERBERT, E. TACKE *et al.*, 2008 Precision breeding for novel starch variants in potato. Plant Biotechnology Journal **6:** 576-584.

MYLES, S., J.-M. CHIA, B. HURWITZ, C. SIMON, G. Y. ZHONG *et al.*, 2010 Rapid Genomic Characterization of the Genus <italic>Vitis</italic>. PloS one **5:** e8219.

MYLES, S., J. PEIFFER, P. J. BROWN, E. S. ERSOZ, Z. ZHANG *et al.*, 2009 Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. The Plant Cell Online **21:** 2194-2202.

NASU, S., J. SUZUKI, R. OHTA, K. HASEGAWA, R. YUI *et al.*, 2002 Search for and Analysis of Single Nucleotide Polymorphisms (SNPs) in Rice (Oryza sativa, Oryza rufipogon) and Establishment of SNP Markers. DNA Research **9:** 163-171.

NATARAJAN, A. T., and G. SHIVASANKAR, 1965 Studies on modification of mutation response of barley seeds to ethyl methanesulfonate. Zeitschrift für Vererbungslehre **96:** 13-21.

NEI, M., and W. H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences **76:** 5269-5273.

NEIGENFIND, J., G. GYETVAI, R. BASEKOW, S. DIEHL, U. ACHENBACH *et al.*, 2008 Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. BMC genomics **9:** 356.

NESTERENKO, S., and K. C. SINK, 2003 Carotenoid Profiles of Potato Breeding Lines and Selected Cultivars. Hortscience **38:** 1173-1177.

NEUFFER, M. G., and E. H. COE, 1978 Paraffin oil technique for treating mature corn pollen with chemical mutagens. Maydica **23:** 8.

NG, P. C., and S. HENIKOFF, 2003 SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research **31:** 3812-3814.

NG, P. C., and S. HENIKOFF, 2006 Predicting the effects of amino acid substitutions on protein function. Annual Review of Genomics and Human Genetics **7:** 61-80.

NG, S. B., E. H. TURNER, P. D. ROBERTSON, S. D. FLYGARE, A. W. BIGHAM *et al.*, 2009 Targeted capture and massively parallel sequencing of 12 human exomes. **461:** 272-276.

NICOT, N., J. F. HAUSMAN, L. HOFFMANN and D. EVERS, 2005 Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. Journal of Experimental Botany **56:** 2907-2914.

NIJMAN, I. J., M. MOKRY, R. VAN BOXTEL, P. TOONEN, E. DE BRUIJN *et al.*, 2010 Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. **7:** 913-915.

NODA, T., S. TSUDA, M. MORI, S. TAKIGAWA, C. MATSUURA-ENDO *et al.*, 2006 Determination of the phosphorus content in potato starch using an energy-dispersive X-ray fluorescence method. Food Chemistry **95:** 632-637.

NORDBORG, M., and D. WEIGEL, 2008 Next-generation genetics in plants. Nature **456:** 720-723.

OBUKHANYCH, T., and J. JURKA, 2007 Rte1_zm. Repbase Rep **7:** 988.

OKOU, D. T., K. M. STEINBERG, C. MIDDLE, D. J. CUTLER, T. J. ALBERT *et al.*, 2007 Microarray-based genomic selection for high-throughput resequencing. **4:** 907-909.

OLIVER, J. L., and J. M. MARTÍNEZ ZAPATER, 1984 Allozyme variability and phylogenetic relationships in the cultivated potato (Solanum tuberosum) and related species. Plant Systematics and Evolution **148:** 1-18.

OLSDER, J., and J. G. T. HERMSEN, 1976 Genetics of self-compatibility in dihaploids of Solanum tuberosum L. I. Breeding behaviour of two self-compatible dihaploids. Euphytica **25:** 597-607.

OHMORI, Y., M. ABIKO, A. HORIBATA and H. Y. HIRANO, 2008 A transposon, Ping, is integrated into intron 4 of the DROOPING LEAF gene of rice, weakly reducing its expression and causing a mild drooping leaf phenotype. Plant and Cell Physiology **49:** 1176-1184.

OTTO, S. P., 2007 The Evolutionary Consequences of Polyploidy. Cell **131:** 452-462.

PARK, T. H., V. G. A. A. VLEESHOUWERS, J. B. KIM, R. C. B. HUTTEN and R. G. F. VISSER, 2005 Dissection of foliage and tuber late blight resistance in mapping populations of potato. Euphytica **143:** 75-83.

PELAK, K., K. V. SHIANNA, D. GE, J. M. MAIA, M. ZHU *et al.*, 2010 The Characterization of Twenty Sequenced Human Genomes. PLoS genetics **6:** e1001111.

POWELL, W., E. BAIRD, N. DUNCAN and R. WAUGH, 1993 Chloroplast DNA variability in old and recently introduced potato cultivars. Annals of Applied Biology **123:** 403-410.

PRITCHARD, J. K., 2001 Deconstructing maize population structure. Nature Genetics **28:** 203-204.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000 Association mapping in structured populations. American Journal of Human Genetics **67:** 170-181.

PROVAN, J., W. POWELL, H. DEWAR, G. BRYAN, G. C. MACHRAY *et al.*, 1999 An extreme cytoplasmic bottleneck in the modern European cultivated potato (Solanum tuberosum) is not reflected in decreased levels of nuclear diversity. Proceedings of the Royal Society of London. Series B: Biological Sciences **266:** 633-639.

PURCELL, S., B. NEALE, K. TODD-BROWN, L. THOMAS, M. A. R. FERREIRA *et al.*, 2007 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics **81:** 559-575.

QUINLAN, A. R., and I. M. HALL, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26:** 841-842.

REED, G. H., and C. T. WITTWER, 2004 Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clinical Chemistry **50:** 1748-1754.

REID, A., L. HOF, D. ESSELINK and B. VOSMAN, 2009 Potato Cultivar Genome Analysis, pp. 295-309 in *Methods in Molecular Biology, Plant Pathology*.

RICKERT, A. M., J. H. KIM, S. MEYER, A. NAGEL, A. BALLVORA *et al.*, 2003 First-generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. Plant Biotechnology Journal **1:** 399-410.

RICKERT, A. M., A. PREMSTALLER, C. GEBHARDT and P. J. OEFNER, 2002 Genotyping of SNPs in a polyploid genome by pyrosequencing (TM). Biotechniques **32:** 592-+.

RITTE, G., M. HEYDENREICH, S. MAHLOW, S. HAEBEL, O. KÃ¶TTING *et al.*, 2006 Phosphorylation of C6- and C3-positions of glucosyl residues in starch is catalysed by distinct dikinases. FEBS Letters **580:** 4872-4876.

RITTE, G., J. R. LLOYD, N. ECKERMANN, A. ROTTMANN, J. KOSSMANN *et al.*, 2002 The starch-related R1 protein is an α-glucan, water dikinase. Proceedings of the National Academy of Sciences of the United States of America **99:** 7166-7171.

RÖMER, S., J. LÜBECK, F. KAUDER, S. STEIGER, C. ADOMAT *et al.*, 2002 Genetic engineering of a zeaxanthin-rich potato by antisense inactivation and co-suppression of carotenoid epoxidation. Metabolic Engineering **4:** 263-272.

ROSS, H., and W. HUNNIUS, 1986 *Potato breeding: problems and perspectives*. P. Parey.

ROUPPE VAN DER VOORT, J. N. A. M., P. VAN ZANDVOORT, H. J. VAN ECK, R. T. FOLKERTSMA, R. C. B. HUTTEN *et al.*, 1997 Use of allele specificity of comigrating AFLP markers to align genetic maps from different potato genotypes. Molecular and General Genetics **255:** 438-447.

SATTARZADEH, A., U. ACHENBACH, J. LÜBECK, J. STRAHWALD, E. TACKE *et al.*, 2006 Single nucleotide polymorphism (SNP) genotyping as basis for developing a PCR-based marker highly diagnostic for potato varieties with high resistance to Globodera pallida pathotype Pa2/3. Molecular Breeding **18:** 301-312.

SCHLÖTTERER, C., and B. HARR, 2002 Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in Drosophila melanogaster. Molecular ecology **11:** 947-950.

SCHORK, N. J., S. S. MURRAY, K. A. FRAZER and E. J. TOPOL, 2009 Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development **19:** 212-219.

SEDDON, J. M., U. A. AJANI, R. D. SPERDUTO, R. HILLER, N. BLAIR *et al.*, 1994 Dietary carotenoids, vitamins A, C, and E, and advanced age-related macular degeneration. Journal of the American Medical Association **272:** 1413-1420.

SEGA, G., 1984 A review of the genetic effects of ethyl methanesulfonate. Mutation Research/Reviews in Genetic Toxicology **134:** 113-142.

SHANNON, P., A. MARKIEL, O. OZIER, N. S. BALIGA, J. T. WANG *et al.*, 2003 Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research **13:** 2498-2504.

SHEN, R., J.-B. FAN, D. CAMPBELL, W. CHANG, J. CHEN *et al.*, 2005 High-throughput SNP genotyping on universal bead arrays. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **573:** 70-82.

SIMKO, I., 2004 One potato, two potato: Haplotype association mapping in autotetraploids. Trends in Plant Science **9:** 441-448.

SIMKO, I., S. COSTANZO, K. G. HAYNES, B. J. CHRIST and R. W. JONES, 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., K. G. HAYNES and R. W. JONES, 2006 Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics **173:** 2237-2245.

SINDEN, S. L., and L. L. SANFORD, 1981 Origin and inheritance of solarmarine glycoalkaloids in commercial potato cultivars. American Potato Journal **58:** 305-325.

SLADE, A. J., S. I. FUERSTENBERG, D. LOEFFLER, M. N. STEINE and D. FACCIOTTI, 2005 A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. **23:** 75-81.

SNODDERLY, D. M., 1995 Evidence for protection against age-related macular degeneration by carotenoids and antioxidant vitamins. American Journal of Clinical Nutrition **62:** 1448S-1461S.

SOMMERBURG, O., J. E. E. KEUNEN, A. C. BIRD and F. J. G. M. VAN KUIJK, 1998 Fruits and vegetables that are sources for lutein and zeaxanthin: The macular pigment in human eyes. British Journal of Ophthalmology **82:** 907-910.

SPOONER, D. M., J. NUÑEZ, F. RODRÍGUEZ, P. S. NAIK and M. GHISLAIN, 2005 Nuclear and chloroplast DNA reassessment of the origin of Indian potato varieties and its implications for the origin of the early European potato. Theoretical and Applied Genetics **110:** 1020-1026.

STADEN, R., 1996 The Staden sequence analysis package. Molecular Biotechnology **5:** 233-241.

STEPHENSON, P., D. BAKER, T. GIRIN, A. PEREZ, S. AMOAH *et al.*, 2010 A rich TILLING resource for studying gene function in Brassica rapa. BMC Plant Biology **10**.

SWINKELS, J. J. M., 1985 Composition and Properties of Commercial Native Starches. Starch - Stärke **37:** 1-5.

SYVANEN, A.-C., 2001 Accessing genetic variation: genotyping single nucleotide polymorphisms. **2:** 930-942.

TAMURA, K., J. DUDLEY, M. NEI and S. KUMAR, 2007 MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular biology and evolution **24:** 1596-1599.

TANAKA, Y., N. SASAKI and A. OHMIYA, 2008 Biosynthesis of plant pigments: Anthocyanins, betalains and carotenoids. Plant Journal **54:** 733-749.

TANG, J., B. VOSMAN, R. E. VOORRIPS, C. G. VAN DER LINDEN and J. A. LEUNISSEN, 2006 QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC bioinformatics **7:** 438.

TENAILLON, M. I., M. C. SAWKINS, A. D. LONG, R. L. GAUT, J. F. DOEBLEY *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proceedings of the National Academy of Sciences **98:** 9161-9166.

TEWHEY, R., M. NAKANO, X. WANG, C. PABÓN-PEÑA, B. NOVAK *et al.*, 2009 Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biology **10**.

THOMSON, M., K. ZHAO, M. WRIGHT, K. MCNALLY, J. REY *et al.*, 2012 High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. Molecular Breeding **29:** 875-886.

THORUP, T. A., B. TANYOLAC, K. D. LIVINGSTONE, S. POPOVSKY, I. PARAN *et al.*, 2000 Candidate gene analysis of organ pigmentation loci in the Solanaceae. Proceedings of the National Academy of Sciences of the United States of America **97:** 11192-11197.

THRALL, P. H., and A. YOUNG, 2000 Autotet: A program for analysis of autotetraploid genotypic data. Journal of Heredity **91:** 348-349.

TILL, B. J., J. COOPER, T. H. TAI, P. COLOWIT, E. A. GREENE *et al.*, 2007 Discovery of chemically induced mutations in rice by TILLING. BMC Plant Biology **7**.

TILL, B. J., S. H. REYNOLDS, E. A. GREENE, C. A. CODOMO, L. C. ENNS *et al.*, 2003 Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Research **13:** 524-530.

TILL, B. J., S. H. REYNOLDS, C. WEIL, N. SPRINGER, C. BURTNER *et al.*, 2004 Discovery of induced point mutations in maize genes by TILLING. BMC Plant Biology **4**.

UIJTEWAAL, B. A., 1987 Ploidy variability in greenhouse cultured and in vitro propagated potato (Solanum tuberosum) monohaploids (2n=x=12) as determined by flow cytometry. Plant Cell Reports **6:** 252-255.

UIJTEWAAL, B. A., D. J. HUIGEN and J. G. T. HERMSEN, 1987 Production of potato monohaploids (2n=x=12) through prickle pollination. Theoretical and Applied Genetics **73:** 751-758.

USUKA, J., W. ZHU and V. BRENDEL, 2000 Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics **16:** 203-211.

VAN BERLOO, R., R. C. B. HUTTEN, H. J. VAN ECK and R. G. F. VISSER, 2007 An online potato pedigree database resource. Potato Research **50:** 45-57.

VAN DER BEEK, J. G., R. VERKERK, P. ZABEL and P. LINDHOUT, 1992 Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: Cf9 (resistance to Cladosporium fulvum) on chromosome 1. Theoretical and Applied Genetics **84:** 106-112.

VAN DE PEER, Y., and R. DE WACHTER, 1994 TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Computer applications in the biosciences : CABIOS **10:** 569-570.

VAN ECK, H. J., J. JACOBS, P. STAM, J. TON, W. J. STIEKEMA *et al.*, 1994 Multiple Alleles for Tuber Shape in Diploid Potato Detected by Qualitative and Quantitative Genetic Analysis Using RFLPs. Genetics **137:** 303-309.

VAN ECK, H. J., and J. E., 1996 Application of molecular markers in the genetic analysis of quantitative traits. Abstract, 13th Triennial Conference of the European Association of Potato Research, EAPR, Veldhoven, The Netherlands**:** 130-131.

VAN OOIJEN, J. W., 2004 MapQTL® 5. Software for the Mapping of Quantitative Trait Loci in Experimental Populations.

VAN OOIJEN, J. W., 2006 JoinMap® 4. Software for the Calculation of Genetic Linkage Maps in Experimental Populations.

VAN ORSOUW, N. J., R. C. J. HOGERS, A. JANSSEN, F. YALCIN, S. SNOEIJERS *et al.*, 2007 Complexity Reduction of Polymorphic Sequences (CRoPS™): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PloS one **2:** e1172.

VAN OS, H., S. ANDRZEJEWSKI, E. BAKKER, I. BARRENA, G. J. BRYAN *et al.*, 2006 Construction of a 10,000-marker ultradense genetic recombination map of potato: Providing a framework for accelerated gene isolation and a genomewide physical map. Genetics **173:** 1075-1087.

VARSHNEY, R. K., S. N. NAYAK, G. D. MAY and S. A. JACKSON, 2009 Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in Biotechnology **27:** 522-530.

VEILLEUX, R. E., and G. J. LIGHTBOURN, 2007 Production and evaluation of somatic hybrids derived from monoploid potato. American Journal of Potato Research **84:** 425-435.

VESELOVSKY, I. A., 1940 Biochemical And Anatomical Properties of Starch of Different Varieties Of Potatoes And Their Importance For Industrial Purposes. American Potato Journal **17:** 330-339.

VISKER, M. H. P. W., L. C. P. KEIZER, H. J. VAN ECK, E. JACOBSEN, L. T. COLON *et al.*, 2003 Can the QTL for late blight resistance on potato chromosome 5 be attributed to foliage maturity type? Theoretical and Applied Genetics **106:** 317-325.

VOORRIPS, R. E., G. GORT and B. VOSMAN, 2011 Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC bioinformatics **12**.

WALDEN, R., 2002 T-DNA tagging in a genomics era. Critical Reviews in Plant Sciences **21:** 143-165.

WANG, Y., A. DIEHL, F. WU, J. VREBALOV, J. GIOVANNONI *et al.*, 2008 Sequencing and comparative analysis of a conserved syntenic segment in the solanaceae. Genetics **180:** 391-408.

WECKX, S., J. DEL-FAVERO, R. RADEMAKERS, L. CLAES, M. CRUTS *et al.*, 2005 novoSNP, a novel computational tool for sequence variation discovery. Genome Research **15:** 436-442.

WERIJ, J., H. FURRER, H. VAN ECK, R. VISSER and C. BACHEM, 2011 A limited set of starch related genes explain several interrelated traits in potato. Euphytica**:** 1-16.

WITTWER, C. T., G. H. REED, C. N. GUNDRY, J. G. VANDERSTEEN and R. J. PRYOR, 2003 High-Resolution Genotyping by Amplicon Melting Analysis Using LCGreen. Clinical Chemistry **49:** 853-860.

WOLTERS, A. M. A., J. G. A. M. L. UITDEWILLIGEN, B. A. KLOOSTERMAN, R. C. B. HUTTEN, R. G. F. VISSER *et al.*, 2010 Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. Plant Molecular Biology **73:** 659-671.

WU, F., L. A. MUELLER, D. CROUZILLAT, V. PETIARD and S. D. TANKSLEY, 2006 Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics **174:** 1407-1420.

XU, X., S. PAN, S. CHENG, B. ZHANG, D. MU *et al.*, 2011 Genome sequence and analysis of the tuber crop potato. Nature **475:** 189-195.

ZEEMAN, S. C., T. DELATTE, G. MESSERLI, M. UMHANG, M. STETTLER *et al.*, 2007 Starch breakdown: Recent discoveries suggest distinct pathways and novel mechanisms. Functional Plant Biology **34:** 465-473.

ZHU, Y. L., Q. J. SONG, D. L. HYTEN, C. P. VAN TASSELL, L. K. MATUKUMALLI *et al.*, 2003 Single-Nucleotide Polymorphisms in Soybean. Genetics **163:** 1123-1134.

Summary

In this thesis natural and induced DNA sequence diversity in potato (*Solanum tuberosum*) for use in marker-trait analysis and potato breeding is assessed. The study addresses the challenges of reliable, high-throughput identification and genotyping of sequence variants in existing tetraploid potato cultivar panels using traditional Sanger sequencing and next-generation massively parallel sequencing (MPS), and the application of this knowledge in the form of genetic markers. Furthermore, it explores the efficiency of ethyl methanesulphonate (EMS) mutagenesis combined with high resolution melting (HRM) DNA screening to induce and discover novel sequence variants in potato genotypes.

Discovery and genotyping of sequence diversity in outcrossing autotetraploid species like potato is complex. In autotetraploid species, genotyping implies the quantitative identification of five alternative allele copy number states. In **Chapter 1**, several methodologies to identify and genotype DNA sequence variants, and the application of these sequence variants is discussed. This chapter provides an introduction to genotyping-by-sequencing (GBS) and the determination of allele copy number.

In **Chapter 2** the sequence diversity in three genes of the carotenoid pathway is assessed in diploid and tetraploid potato genotypes using direct Sanger sequencing. To investigate the genetics and molecular biology of orange and yellow flesh colour in potato, association analysis between SNP haplotypes and flesh colour phenotypes was performed, and the inheritance and gene expression of associated alleles was studied. We observed among eleven beta-carotene hydroxylase 2 (*CHY2*) alleles one dominant allele with a major effect, changing white into yellow flesh colour. In contrast, none of the lycopene epsilon cyclase (LCYe) alleles seemed to have a large effect on flesh colour. Analysis of zeaxanthin epoxidase (ZEP) alleles showed that a recessive allele with a non-LTR retrotransposon sequence in intron 1 reduced the expression level of the ZEP gene and caused accumulation of zeaxanthin. Genotypes combining presence of the dominant CHY2 allele with homozygosity for the recessive ZEP allele produced orange-fleshed tubers that accumulate large amounts of zeaxanthin.

Sanger amplicon sequencing was applied in **Chapter 3** to evaluate the sequence diversity in α-Glucan Water Dikinase (*StGWD*), a candidate gene underlying a QTL involved in potato starch phosphate content. Sanger sequences of two *StGWD* amplicons from a global collection of 398 commercial cultivars and progenitor lines were used to identify 16 unique haplotypes. By assigning tag SNPs to these haplotypes and by determining the allele copy number of identified sequence variants, we inferred the four-allele genetic composition for almost all cultivars assayed at this locus. This allowed genetic diversity parameters like the average number of different alleles present in a single cultivar (*Ai*=3.1) and the average intra-individual heterozygosity (*Ho*=0.765) to be estimated for this locus. Pedigree analysis confirmed that the identified haplotypes are identical by descent (IBD) and offered insight in the breeding history of elite potato germplasm. Haplotype association analysis led to the identification of two *StGWD* alleles causing altered starch phosphate content, which was further verified in diploid and tetraploid mapping populations containing the relevant alleles.

One of these alleles (Allel H) increases the fraction of starch that is phosphorylated, while the other one (Allele A) decreases it.

To scale up the discovery and genotyping of sequence variants, and to make it more whole-genome oriented, **Chapter 4** reports on massively parallel sequencing (MPS) of approximately 800 genes scattered over the potato genome and resequenced in 83 tetraploid potato cultivars and a monoploid reference accession. We show that by combining MPS with genome complexity reduction and indexed sequencing, sufficient read depth for GBS can be achieved for reliable discovery and genotyping of sequence variants in individual tetraploid potato genotypes. With a custom designed, SureSelect enrichment library, 1.44 Mb of DNA sequence was targeted. The genes targeted were mainly single-copy genes, selected based on putative gene functions in both primary and secondary metabolic pathways, potato quality traits and biotic and abiotic stresses, and included a large set of conserved orthologous sequence genes (COSII) useful for genetic anchoring and phylogenetic studies. The indexed and enriched DNA libraries were sequenced on a Illumina HiSeq. After filtering and processing the raw sequence data, 12.4 Gb of high-quality sequence data was mapped to the potato genome, covering 2.1 Mb of the genome sequence with a median average read depth of 63× per cultivar. We detected over 129,000 sequence variants in these data and determined allele copy number of the variants in individual potato samples. The accuracy of the sequence-based allele copy number estimates was verified by a low-density SNP genotyping assay. This showed that for reliable genotyping a read count-based genotype quality score is best applied and a read depth of 80× is recommended for determining allele copy number in autotetraploid potato. Average nucleotide diversity (π=10.7×10^{-3} genome-wide, ≈1 variant/93 bp between two random alleles) varied along the twelve potato chromosomes, and individual genes under selection were identified. As an example for application of GBS for genome-wide association analysis (GWAS), the identified sequence variants and genotype data were tested in a marker-trait association analysis with plant maturity and tuber flesh colour. This led to the identification of alleles accounting for significant phenotypic variation in these traits.

In **Chapter 5** we applied the chemical mutagen EMS to diploid potato by two different treatments, a pollen and a seed treatment. We screened the resulting populations for novel mutations using HRM analysis. A pollen treatment with EMS dissolved in a sucrose solution was found to induce mutations only at a low frequency (only one mutation discovered after screening >2.7 Mb of sequence). *In planta* selection of the most vital mutagenized pollen seems to have lowered the mutation density to a frequency that is not suitable for reverse genetics studies. Treatment of potato seeds with EMS on the other hand provided a high density of novel mutations (1 mutation/65 kb), discovered in the M_1 generation. In contrast to most EMS mutagenesis studies, we directly screened the M_1 generation of the seed-treated population. In six candidate genes involved in potato starch and frying quality traits, 65 novel sequence variants were discovered. In all six genes, missense mutations that are predicted to damage protein function were discovered, and for four genes five premature stop codon mutations were identified. We attempted to stabilize and transfer 27 putatively interesting mutations to the M_2 and M_3 generation for further evaluation. Genetically stable M_2 and M_3 plants have been generated for 10 (37%) of these mutations. The estimated density of M_1 mutations that

are transferable to the M_2 generation (one "accessible" mutation/118-176 kb) is higher than the mutation densities obtained in most other plant species, for which the M_2 generation has been screened. The results of this chapter thus demonstrate that screening the M_1 generation offers a good alternative to the commonly applied M_2 screening for the rapid generation of novel genetic variation at a high density, without too much complication in recovering mutations in the M_2 generation.

In the concluding **Chapter 6**, results of preceding chapters are evaluated, and the prospects of the findings for potato research and breeding are discussed.

Samenvatting

In dit proefschrift wordt de natuurlijke en nieuw geïnduceerde DNA-sequentie diversiteit in aardappel (*Solanum tuberosum*) gebruikt om de relatie tussen genotype en fenotype te bestuderen, en deze kennis toe te passen in de vorm van moleculaire merkers voor gebruik in de aardappelveredeling. De studie richt zich op de betrouwbare identificatie en genotypering van sequentievarianten in bestaande tetraploïde aardappelrassen met behulp van twee methoden van DNA-sequentiebepaling (de conventionele Sanger methode en een methode van de volgende generatie met grootschalige verwerkingscapaciteit). Verder wordt, om nieuwe sequentievariatie te induceren en identificeren, de efficiëntie van ethyl methaansulfonaat (EMS) mutagenese en de detectie van geïnduceerde sequentievarianten met behulp van hoge resolutie smeltanalyse (HRM) onderzocht.

Het identificeren en de genotypering van DNA sequentie diversiteit in autotetraploïde uitkruisende gewassen zoals aardappel is complex. In autotetraploïde soorten zijn er per allel vijf verschillende kopie-aantallen per genotype mogelijk, en voor genotypering moeten deze alternatieve situaties kwantitatief vastgesteld kunnen worden. In **Hoofdstuk 1** worden verschillende methoden voor het identificeren van DNA sequentievarianten, genotypering en de toepassing van DNA sequentievariatie besproken. Dit hoofdstuk geeft een introductie tot genotypering-door-sequentiebepaling (GBS) en de bepaling van het kopie-aantal van allelen.

In **Hoofdstuk 2** wordt de DNA sequentie diversiteit in drie genen van de carotenoïd biosynthese route bestudeerd in diploïde en tetraploïde aardappels met behulp van de Sanger-sequentiebepalin, direct toegepast op PCR amplificatieproducten die een mengsel van allelen omvat. Om de genetica en moleculaire biologie van oranje en gele vleeskleur in aardappel te onderzoeken werd associatie analyse tussen SNP haplotypes en vleeskleur fenotypes uitgevoerd en werd de overerving en genexpressie van de bijbehorende allelen bestudeerd. Van de elf beta-caroteen hydroxylase 2 (CHY2) allelen identificeerden we één allel met een dominant effect, dat de kleur van het vruchtvlees veranderde van wit naar geel. Géén van de lycopeen epsilon cyclase (LCYe) allelen leek een groot effect op vleeskleur hebben. Analyse van de zeaxanthine epoxidase (ZEP) allelen toonde aan dat een recessief allel met een niet-LTR retrotransposon sequentie in intron 1 het expressieniveau van het ZEP gen vermindert en accumulatie van zeaxanthine veroorzaakt. Genotypen die de aanwezigheid van het dominante CHY2 allel combineren met homozygotie voor het recessieve ZEP allel produceren aardappelknollen met oranje vleeskleur die grote hoeveelheden zeaxanthine accumuleren.

Sanger amplicon sequentiebepaling wordt ook toegepast in **Hoofdstuk 3** om de sequentie diversiteit in het gen α-glucan Water Dikinase (*StGWD*) te evalueren, een kandidaatgen dat ten grondslag ligt aan een QTL voor het fosfaatgehalte in aardappelzetmeel. Door analyse van Sanger sequenties van twee *StGWD* amplicons van een mondiaal representatieve verzameling van commerciële rassen en ouderlijnen konden 16 unieke *StGWD* haplotypes geïdentificeerd worden. Door het toewijzen van tag SNPs aan deze haplotypes en door bepaling van het kopie aantal van alle geïdentificeerde sequentievarianten en allelen kon de volledige genetische samenstelling op dit locus voor bijna alle geanalyseerde cultivars afgeleid worden. Hierdoor

konden voor dit locus genetische diversiteit parameters zoals het gemiddelde aantal verschillende allelen die aanwezig zijn in een ras (Ai = 3,1) en de gemiddelde intra-individuele heterozygotie (Ho = 0,765) berekend worden. Afstammingsanalyse bevestigde dat de geïdentificeerde haplotypes identiek zijn door afstamming (IBD) en dit bood inzicht in de veredelingsgeschiedenis van aardappel. Haplotype associatie analyse leidde tot de identificatie van twee *StGWD* allelen die een veranderd fosfaatgehalte in zetmeel veroorzaken. Het effect van deze allelen werd verder geverifieerd in diploïde en tetraploïde kruisingspopulaties met de relevante allelen. Een van deze allelen (Allel *H*) verhoogt de fosforyleringsgraad van aardappelzetmeel, terwijl het andere (Allel *A*) het verlaagt.

Voor grootschalige identificatie en genotypering van DNA sequentievarianten, beschrijft **Hoofdstuk 4** de massaal parallelle sequentiebepaling (MPS) van ongeveer 800 verspreid gelegen genen uit het aardappelgenoom. De sequentievariatie van deze genen werd bepaald in 83 tetraploïde aardappelrassen en een monoploïde referentie. Door MPS te combineren met enerzijds een geïndexeerde sequentiebepaling en anderzijds een genoom complexiteit reductie kon voldoende sequentiediepte worden bereikt voor een betrouwbare identificatie en genotypering van sequentievarianten in de individuele genotypen. De genoomcomplexiteitsreductie werd gerealiseerd met een voor dit doel ontworpen SureSelect genoom verrijkingsbibliotheek met 1,44 Mb aan beoogde DNA-sequentie. Beoogde doelwit genen waren voornamelijk genen geselecteerd op basis van vermeende genfuncties in zowel de primaire als secundaire metabolische routes, aardappel kwaliteitseigenschappen en biotische en abiotische stress resistenties en omvatte een groot aantal genen met een geconserveerde orthologe sequentie (COSII). Deze laatste zijn handig voor verankering op een genetische kaart en voor fylogenetische studies. Sequentiebepaling van de geïndexeerde en verrijkte DNA collecties werd uitgevoerd met een Illumina HiSeq apparaat. Na schifting en verwerking van de ruwe sequentie data werd 12,4 Gb aan hoogwaardige sequentiegegevens toegekend, uitgelijnd en geprojecteerd op 2,1 Mb van het aardappelgenoom, met een gemiddelde sequentiediepte van 63× per cultivar. In deze data werden meer dan 129.000 sequentievarianten ontdekt en het kopie-aantal van de varianten werd in de afzonderlijke cultivars bepaald. De nauwkeurigheid van deze op sequentie gebaseerde allelschattingen werd bevestigd met behulp van een onafhankelijke SNP genotyperingsassay. Hieruit bleek dat voor een betrouwbare genotypering op basis van sequentiediepte een genotype kwaliteitscore moet worden toegepast en dat een sequentiediepte van 80× geschikt is voor het bepalen van het kopie-aantal van allelen in autotetraploïde gewassen zoals aardappel. De gemiddelde nucleotide diversiteit (genoom breed $\pi = 10{,}7 \times 10^{-3}$, ≈ 1 variant/93 bp tussen twee willekeurige allelen) varieerde tussen de twaalf aardappel chromosomen, en individuele genen die onder selectie staan werden geïdentificeerd. Als voorbeeld voor de toepassing van GBS voor genoom-brede associatie analyse (GWAS) werden de geïdentificeerde sequentievarianten en genotype gegevens in verband gebracht met kenmerken zoals vroegrijpheid en vleeskleur in een DNA merker-kenmerk associatie analyse. Dit leidde tot de identificatie van allelen die een belangrijk aandeel leveren aan de fenotypische variatie voor deze eigenschappen.

In **Hoofdstuk 5** hebben we het chemische mutagen EMS toegepast op diploïd aardappelmateriaal in twee verschillende behandelingen; een pollenbehandeling en een

zaadbehandeling. De resulterende nakomelingschappen werden geanalyseerd op hun DNA smeltpunt (HRM analyse) om zodoende nieuw-geïnduceerde mutaties te detecteren. De behandeling van pollen met EMS opgelost in een sucrose-oplossing bleek slechts met een lage frequentie mutaties te induceren (slechts één mutatie ontdekt na > 2,7 Mb DNA sequentie screening). In planta selectie van het meest vitale gemutageniseerde pollen lijkt de mutatie dichtheid te hebben gereduceerd tot een frequentie die niet geschikt is voor "reverse genetics" studies. Behandeling van zaden met EMS voorzag wel in een hoge dichtheid van geïnduceerde mutaties (1 nieuwe mutatie/65 kb in de M_1 generatie). In tegenstelling tot de meeste EMS-mutagenese studies werd in dit hoofdstuk de M_1 generatie van de zaad-behandelde populatie direct gescreend. In zes kandidaatgenen die betrokken zijn bij aardappelzetmeel kenmerken en bakkwaliteit, werden 65 geïnduceerde sequentie-varianten ontdekt. In alle zes de genen werden missense mutaties ontdekt waarvan wordt aangenomen dat ze schade aan de eiwitfunctie veroorzaken en in vier genen werden vijf voortijdige stopcodon mutaties geïdentificeerd. Voor genetische stabilisatie en verdere evaluatie werd voor 27 van de meest interessante mutaties geprobeerd om de mutaties over te brengen naar de M_2 en M_3 generaties. Genetisch stabiele M_2 en M_3 planten zijn verkregen voor 10 (37%) van deze mutaties. De geschatte dichtheid van M_1 mutaties die overdraagbaar zijn naar de M_2 generatie (een "toegankelijke" mutatie per 118-176 kb) is hoger dan de mutatie dichtheid die voor de meeste andere plantensoorten wordt verkregen en waarbij de M_2 generatie wordt onderzocht. De resultaten van dit hoofdstuk tonen aldus aan dat het screenen van de M_1 generatie een goed alternatief is voor de gangbare M_2 screening en toegepast kan worden om snel nieuwe genetische variatie te genereren, met een hoge mutatiedichtheid en zonder te veel complicaties in het terugvinden van de geïdentificeerde mutaties in de M_2 generatie.

In het afsluitende **Hoofdstuk 6** worden de resultaten van de voorgaande hoofdstukken geëvalueerd en de vooruitzichten van de bevindingen voor aardappelonderzoek en -veredeling besproken.

Dankwoord

Een woord van dank aan iedereen die op enige wijze heeft bijgedragen aan de totstandkoming van dit proefschrift. In het bijzonder de docenten en begeleiders die belangrijk zijn geweest tijdens mijn eerdere opleidingen; Truus Rigter, Jules Beekwilder, Bertrand Gakière en Bjorn Kloosterman. De directe begeleiders betrokken bij dit proefschrift; Herman van Eck, Anne-Marie Wolters en Richard Visser. De leden van de STW gebruikerscommissie en deelnemende bedrijven; AVEBE/Averis Seeds, Aviko, C. Meijer, McCain Foods Holland, AGRICO Research, HZPC Holland, Van Rijn/KWS. Paranimfen Peter Vos en Richard Remeeus, collega's, vrienden en familie. Pap en mam bedankt voor de ruimte in een brede omgeving. Sjir en Janneke voor de competitieve drive in een ontspannen omgeving. Dion voor het groots denken in een vakantie omgeving. Lianne's familie voor de tolerantie in een soms gespannen omgeving. Lianne voor liefde en een warme omgeving.

Over de Auteur

Jan Uitdewilligen werd op 17 maart 1979 geboren in Thull, Schinnen. In 1996 behaalde hij aan het St. Jans College te Hoensbroek het MAVO-diploma. In datzelfde jaar begon zijn middelbaar laboratorium onderwijs aan het Leeuwenborg te Sittard. In 2001 behaalde hij zijn diploma en begon hij zijn studie Biotechnologie aan het van Hall Instituut / Noordelijke hogeschool Leeuwarden met als specialisatierichting plantenbiotechnologie. Zijn stage en afstudeeropdracht heeft hij verricht bij Plant Research International en het Max Planck Instituut voor Plantenfysiologie. Na in 2005 zijn diploma behaald te hebben begon hij zijn studie plantenveredeling en genetische bronnen aan de Wageningen Universiteit. Het afstudeeronderzoek voor deze studie voerde hij uit bij de vakgroep Plantenveredeling. In 2007 verkreeg hij zijn diploma met als afstudeerrichting moleculaire plantenveredeling. In december van datzelfde jaar begon hij als AIO bij het laboratorium voor Plantenveredeling aan de Wageningen Universiteit. Het promotieonderzoek richtte zich op de identificatie en genotypering van DNA sequentie variatie in aardappel en de koppeling van deze variatie aan fenotypische kenmerken, waarvan de resultaten beschreven staan in dit proefschrift.

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: Jan Uitdewilligen
Date: 18 September 2012
Group: Laboratory of Plant Breeding, Wageningen University & Research Centre

1) Start-up phase	*date*
► **First presentation of your project**	
"Potatoes with novel properties for consumption and processing industry"	21 Dec 2007
► **Writing or rewriting a project proposal**	
► **Writing a review or book chapter**	
► **MSc courses**	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*1,5 credits**

2) Scientific Exposure	*date*
► **EPS PhD Student Days**	
1st joined retreat of PhD students in experminental Plant Sciences, Wageningen	02-03 Oct 2008
EPS PhD student day, Utrecht	01 Jun 2010
EPS Career day, Wageningen	18 Nov 2010
EPS PhD student day, Wageningen	20 May 2011
► **EPS Theme Symposia**	
EPS Theme Symposium 4 'Genome Plasticity', Wageningen	12 Dec 2008
EPS Theme Symposium 4 'Genome Plasticity', Nijmegen	11 Dec 2009
EPS Theme Symposium 4 'Genome Plasticity', Wageningen	10 Dec 2010
EPS Theme Symposium 4 'Genome Plasticity', Wageningen	09 Dec 2011
► **NWO Lunteren days and other National Platforms**	
NWO-ALW Experimental Plant Sciences, Lunteren	07-08 Apr 2008
NWO-ALW Experimental Plant Sciences, Lunteren	06-07 Apr 2009
NWO-ALW Experimental Plant Sciences, Lunteren	19-20 Apr 2010
NWO-ALW Experimental Plant Sciences, Lunteren	02-03 Apr 2011
► **Seminars (series), workshops and symposia**	
Linkage disequilibrium and association mapping – helping to overcome the paradox of modern plant breeding' by Wallace Cowling	26 Jun 2009
STW jaarcongress	08 Oct 2009
The molecular basis of quantitative traits in potato' by Christiane Gebhardt	05 Feb 2010
'Mobile RNA silencing in plants' by David Baulcombe	27 Sep 2010
The SOL Genomics Network: Genome databases in a post-genome world' by Lukas Mueller	04 Oct 2010
Agilent Microarray en Sequencing Roadshow	02 Jun 2010
Bioinformatics@PBR	02 Feb 2011
Agilent Genomics & Automation seminar	01 Mar 2011
EPS Mini-symposium 'Plant Breeding in the genomics era'	25 Nov 2011
Metabolic engineering of high-value industrial and nutritional isoprenoids in plants' by Dr. Paul Fraser	16 Feb 2012
The Tomato Genome: From Genes To QTL and Networks' by Graham Seymour	24 Jan 2012
Plant Breeding Research days	2009-2012
WUR Plant Sciences Seminars	2010-2012
Studiekring Plantenveredeling	2009-2012
► **Seminar plus**	
► **International symposia and congresses**	
The 5th Solanaceae Genome Workshop (SOL2008)	12-16 Oct 2008
The 6th Solanaceae Genome Workshop (SOL2009)	08-13 Nov 2009
EAPR-EUCARPIA congress "Potato Breeding after completion of the DNA Sequence of the Potato Genome"	27-30 Jun 2010
STATSEQ WG2 Genotyping by Sequencing	03-04 Nov 2011
► **Presentations**	
Oral presentation at EPS PhD workshop 'Natural Variation in Plants'	28 Sep 2008
Oral presentation at the 5th Solanaceae Genome Workshop (SOL2008)	15 Oct 2008
Oral presentation at Biometris StatGen colloquium	03 Mar 2009
Poster at the 6th Solanaceae Genome Workshop (SOL2009)	08-13 Nov 2009
Oral presentation at EAPR-EUCARPIA 2010	30 Jun 2010
Poster at Plant Genome Evolution 2011	04-06 Sep 2011
Oral presentation at STATSEQ WG2 Genotyping by Sequencing	03 Nov 2011
Oral presentation at EPS Theme Symposium 4 'Genome Plasticity'	09 Dec 2011
► **IAB interview**	04 Dec 2009
► **Excursions**	
KeyGene; McCain Nederland	27 Aug & 15 Sep 2008
Meijer Potato Breeding	14 Jul 2009
Averis Seeds	15 Jul 2010
EPS excursion Monsanto; Van Haeringen Laboratorium; PCDI Green Life Science Company Visists	27 Jan, 17 & 23 Jun 2011
Subtotal Scientific Exposure	*23,8 credits**

3) In-Depth Studies	*date*
► **EPS courses or other PhD courses**	
WIAS course Statistics for the Life Sciences	28 May-04 Jun 2008
Summer School 'On the Evolution of Plant Pathogen Interactions: from Principles to Practice'	18-20 Jun 2008
EPS PhD workshop 'Natural Variation in Plants'	26-29 Aug 2008
Next Generation Sequencing Course LUMC leiden	29 Aug-01 Sep 2010
► **Journal club**	
Member of literature discussion group of Plant Breeding	2008-2011
► **Individual research training**	
Illumina HiSeq sequencing by Pieter van der Vlies, UMCG; DNA sample preparation by Alexanter Hoischen, UMC Radboud Nijmegen	Apr 2010
Subtotal In-Depth Studies	*8,4 credits**

4) Personal development	*date*
► **Skill training courses**	
PhD Competence Assessment	May-Jun 2008
Course: Scientific writing	Feb-April 2009
Course: Career perspectives	Mar-Apr 2011
► **Organisation of PhD students day, course or conference**	
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*3,9 credits**

TOTAL NUMBER OF CREDIT POINTS*	**37.6**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

This research was financially supported by
the Dutch Technology Foundation STW (WPB-7926)

Thesis layout and cover design by the author

Printed by Wöhrman Print Service, Postbus 92, 7200 AB, Zutphen, NL