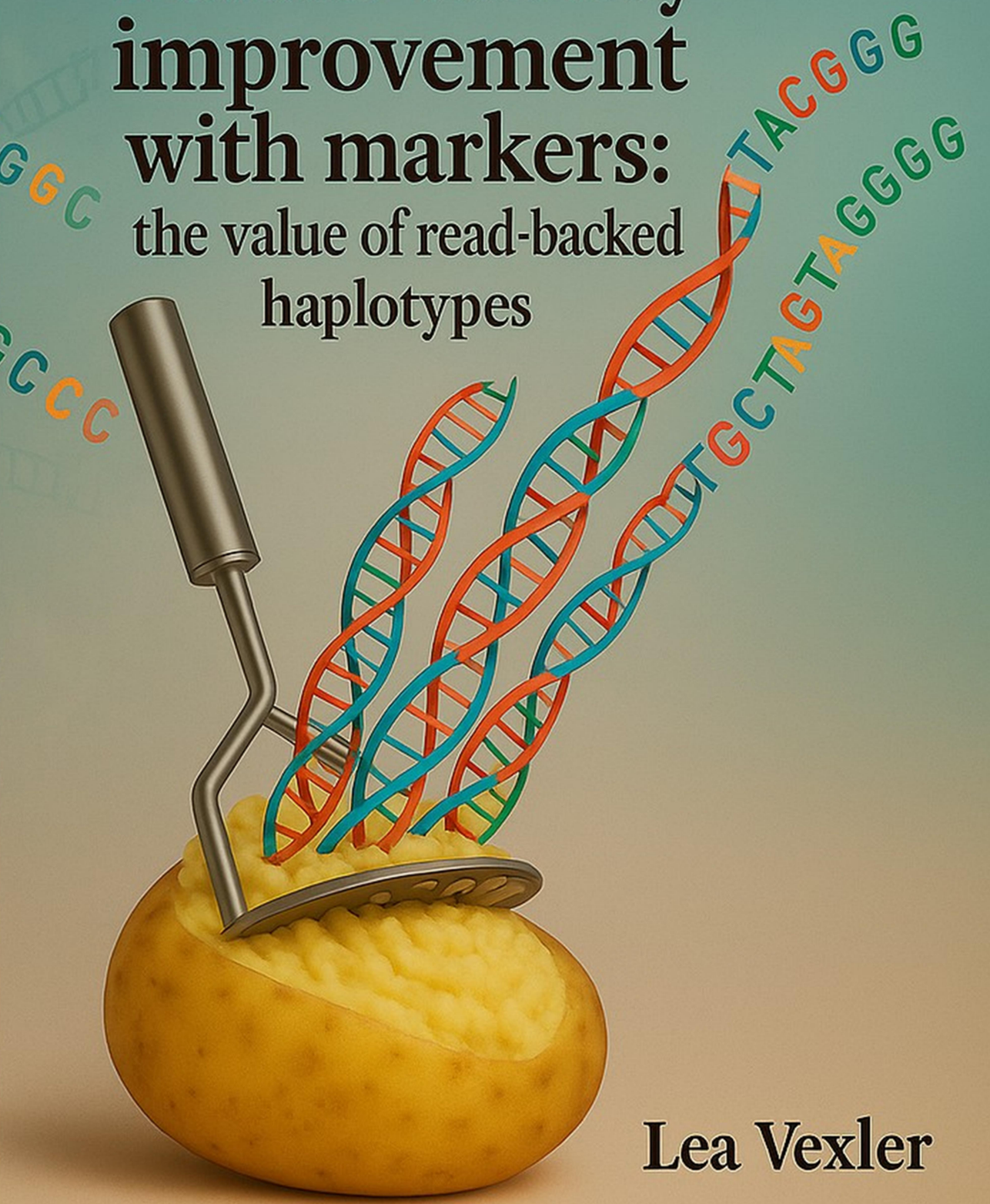
Potato variety improvement with markers:

the value of read-backed haplotypes



Lea Vexler

Propositions

1. One thesis lowers the cost of an entire breeding program.
   (this thesis)

2. Fixation and heterozygosity are not opposites - they are allies.
   (this thesis)

3. Like in the sport of weightlifting, science rewards consistency over ego.

4. The success of human genetic therapies depends as much on psychology and sociology as it does on molecular biology.

5. Genomics can predict about 50% of athletic performance potential, the real scientific challenge is uncovering the rest.

6. Restoring ecological function with proxy species is more important than recreating genetic authenticity in rewilding strategies.

7. Diversity is a trend for institutions; true inclusion is earned by individuals.

Propositions belonging to the thesis, entitled:

“Potato variety improvement with markers: the value of read-backed haplotypes”

Lea Vexler

Wageningen, 26 September 2025

Potato improvement with markers: the value of read-backed haplotypes

Lea Vexler

Thesis committee

Promotor

Prof. dr R.G.F. Visser
Professor in Plant Breeding
Wageningen University & Research

Co-promotors

Dr Ir. H.J. van Eck
Assistant professor, Plant Breeding
Wageningen University & Research

Dr D. Milbourne
Senior Research officer
Crop Science Department, Teagasc, Ireland

Other members

Prof. dr F.A. van Eeuwijk, Wageningen University & Research

Dr M.W.M. Muskens, Agrico Research, The Netherlands

Dr H.H. Tai, Fredericton Research and Development Centre, Agriculture and Agri-Food, Canada

Dr P.G. Vos, HZPC, The Netherlands

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

Potato improvement with markers: the value of read-backed haplotypes

Lea Vexler

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. dr C. Kroeze
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 26 September 2025
at 3.30 p.m. in the Omnia Auditorium

Lea Vexler
Potato improvement with markers: the value of read-backed haplotypes,
216 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2025)
With references, with summary in English

ISBN: 978-94-6510-787-5

DOI: 10.18174/681248

In loving memory of Sharon Brennan,
who brought love, care, and light to the lives around her.
Your kindness lives on.

Table of contents

Chapter 1 **9**

General Introduction

Chapter 2 **29**

PotatoMASH - A Low Cost, Genome-Scanning Marker System for Use in Potato Genomics and Genetics Applications

Chapter 3 **61**

QTL discovery for agronomic and quality traits in diploid potato clones using PotatoMASH amplicon sequencing.

Chapter 4 **97**

Utilizing Multiplex Amplicon Sequencing and Read-Backed Haplotyping to Track Homozygosity and Residual Heterozygosity in Diploid Potato Breeding

Chapter 5 **123**

PotatoMASH is a cost-effective marker system for Genomic Prediction in potato based on short-read haplotypes

Chapter 6 **161**

General Discussion

References **185**

Summary **203**

Acknowledgments **209**

About the Author **213**

Authorship Statement **215**

Chapter 1

General Introduction

Evolving strategies in potato breeding

Potato (*Solanum tuberosum*) is one of the most important staple crops globally, contributing substantially to global food security and agricultural economies (FAO Crops statistics database: http://faostat.fao.org/). Given its global importance, the ability to rapidly and precisely breed potato varieties that combine multiple favourable traits has been widely recognized as a priority. Breeders must integrate as approximately 40 important traits (Gebhardt 2013; Slater et al. 2016), including various disease resistances, tolerance to abiotic stresses, yield stability, agronomic and tuber quality traits, making the breeding process highly complex and resource-intensive.

Conventional potato breeding occurs at the tetraploid level (2n = 4x = 48) to exploit the yield advantage conferred by larger cell size (Mendiburu and Peloquin 1977). The improvement of potato varieties primarily relies on clonal selection following the hybrid crosses between highly heterozygous tetraploid parents, each possessing four sets of 12 chromosomes. Meiosis in tetraploids involves genetic recombination events, leading to an astronomical number of possible gamete combinations (Figure 1a). This requires large population sizes to effectively select for desirable traits (Bradshaw 2007; Ortiz and Mihovilovich 2020). Despite intensive selection, desirable allele combinations of a superior individual will disperse during the next meiosis. In this thesis, we use the terms fixed or fixation to indicate the homozygous state of certain loci, with the purpose that all offspring will consistently inherit superior parental alleles. Without fixation, it is difficult for breeders to reach the final ideotype, and the offspring descending from non-inbred parents is highly unpredictable. Over decades, breeders have prioritized high heterozygosity to enhance yield and hybrid vigour (heterosis). While this strategy has been beneficial for maintaining genetic diversity, it has also led to the accumulation of deleterious alleles in tetraploid potatoes. These unfavourable alleles remain "hidden" within the genome and “re-emerge” in each breeding cycle, complicating genetic improvement (Bradshaw 2017). Large-scale selection programs are necessary to achieve a balance between unfavourable and beneficial alleles at different loci.

Currently, phenotypic selection across clonal generations remains the primary approach for improving potato. Typically, breeding begins with about 100,000 seedlings from 200–300 crosses, followed by clonal selection over several years (Ghislain and Douches 2020; Bradshaw 2017). The scale and duration of this process are exemplified by the Teagasc-IPM Potato Breeding Programme, illustrated in Figure 1b. This programme follows the widely used "early generation intensive selection" model, which operates within a structured 12-year cycle of trials and selections. It begins with initial crosses in Year 1, raising seedlings in Year 2, followed by intense selection pressure in Year 3, during the first field stage, where poor-performing seed tuber-grown genotypes are rapidly discarded. This enables more focused selection on a reduced subset of promising individuals. As the programme progresses, selection pressure decreases, while the scale and complexity of evaluations increase through multi-location and replicated field trials. In addition to this conventional

breeding pipeline, the programme also incorporates a Rapid Cycle breeding scheme (see Figure 1b), aimed at accelerating the development of varieties with stacked resistance genes. This approach is discussed further in the section on genomic-assisted breeding.

Efforts to improve the efficiency and predictability of potato breeding have evolved significantly over time, reflecting a growing understanding of the crop's genetic complexity. One of the earliest proposed strategies was Chase's analytical breeding approach (1963), which outlined a three-step process: 1) reducing ploidy, 2) breeding at the diploid level, and 3) re-synthesizing tetraploids. This framework aimed to exploit the simpler genetics of diploids while ultimately returning to the high-yielding tetraploid level for commercial cultivation.

A major breakthrough followed in the 1970s–1990s, when Peloquin and colleagues demonstrated the practical application of unreduced (2n) gametes in potato breeding. Their work on first division restitution (FDR) 2n gametes showed that approximately 80% of the diploid genetic complement could be transmitted to tetraploid offspring, preserving heterozygosity and enhancing the transmission of desirable traits (Hutten et al. 1994; Kidane-Mariam and Peloquin 1975; Mendiburu and Peloquin 1971; Mendiburu and Peloquin 1977; Ortiz et al. 1991; Werner and Peloquin 1990). Around this time, the International Potato Center (CIP) attempted to utilize 4x × 2x breeding to develop true potato seed (TPS), aiming for a higher reproduction rate and to reduce the logistical burden of distributing bulky seed tubers (Jackson 1987). However, TPS-based systems did not gain widespread adoption due to limited agronomic knowledge on TPS handling and propagation, as well as poor uniformity in progenies, which negatively affected agronomic consistency (Clot 2023).

In the early 2000s, Bradshaw et al. (2003) proposed a multi-trait selection scheme to accelerate recurrent selection in tetraploid breeding. This approach combined mid-parent values (parental trait averages), early progeny testing, and a selection index covering key traits such as yield, disease resistance, and processing quality. The scheme shortened the breeding cycle from nine to three years by eliminating entire progenies from poor parent combinations early and applying direct trait assessments to the remaining clones. However, despite its potential to improve both genetic gain and breeding efficiency compared to traditional clonal selection, commercial programs did not adopt this strategy, likely due to the continued complexity of tetraploid genetics and the operational demands of implementing family-based selection schemes.

Although the potential of diploid breeding was recognized as early as the 1950s (Hougas and Peloquin 1958), it gained renewed attention with the introduction of the F1 hybrid breeding system (Lindhout et al. 2011). The system relies on developing diploid homozygous inbred lines that can be crossed to produce vigorous, uniform F1 seed. However, progress in this area requires to overcome two key biological constraints: self-incompatibility and inbreeding depression. Self-incompatibility in diploid potatoes, which prevents self-fertilization, has long been a major obstacle (De Jong and Rowe 1971b; Olsder and Hermsen 1976; Phumichai and Hosaka 2006). A significant breakthrough came when

Hosaka and Hanneman (1998a, 1998b) identified the *Sli* locus on chromosome 12, which enables self-compatibility. More recently, Clot et al. (2020) showed that the *Sli* haplotype is widespread across both tetraploid and diploid cultivated germplasm, and developed diagnostic genomic markers that have been rapidly adopted for marker-assisted selection (Kaiser et al. 2020; Song and Endelman 2023; Sood et al. 2022).

Despite overcoming self-incompatibility, inbreeding depression remains a major limitation. Selfing exposes recessive deleterious alleles, leading to reduced vigour, poor growth, and fertility issues such as flower abortion and pollen sterility (Charlesworth and Willis 2009; De Jong and Rowe 1971b; Hosaka and Sanetomo 2020; Krantz 1924; Krantz 1946; Peterson et al. 2016; Phumichai and Hosaka 2006; Wu et al. 2023; Zhang et al. 2019). In recent years, F1 hybrid breeding has gained momentum globally, with several programs focusing on diploid systems using self-compatible lines (Bradshaw 2022; Hosaka and Sanetomo 2020; Jansky et al. 2016; Zhang et al. 2022b). In this approach, highly homozygous parental lines are crossed to produce uniform F1 botanical seed, with the negative effects of inbreeding mitigated through heterosis (De Jong and Rowe 1971b; Lindhout et al. 2011). However, the development of fully homozygous inbred lines remains a substantial challenge, and it is still uncertain whether on average diploid hybrids, even with the advantage of heterosis, can match the high yield potential of elite tetraploid cultivars.

Clot in his PhD thesis (2023) presented a novel strategy called Fixation-Restitution Breeding (Fix-Res) (Figure 2). This approach captures many of the benefits of diploid hybrid potato breeding and avoids its disadvantages, while enabling the results of diploid breeding to be transferred into the tetraploid background through interploidy crosses. In the first step, similar to diploid F1 hybrid breeding, the diploid potato is rendered self-compatible by introgression of the monogenic dominant self-compatibility gene (*Sli*), allowing inbreeding and recurrent backcrossing to accumulate and fix beneficial alleles. Selfing is only employed to reach fixation of beneficial trait alleles, but not necessarily to increase homozygosity. Unlike traditional diploid F1 hybrid breeding, the diploid potato's ability to produce diploid gametes through meiotic chromosome restitution mechanisms is utilized in the second step. This chromosome restitution step combined with the fixation step are the two elementary features of the Fixation-Restitution breeding scheme. Interploidy crossing between the diploid and a tetraploid ensure a near complete transfer of the chromosomal content from the diploid to the tetraploid. This feature is a significant advantage of Fix-Res Breeding, as it ensures the preservation of beneficial allele combinations from the diploid parent, even if they are in a heterozygous state. By maintaining higher heterozygosity, Fix-Res integrates controlled inbreeding at strategic loci while avoiding the loss of vigour and fertility due to inbreeding.

Potato breeding has long been constrained by its outbreeding nature and tetraploid complexity, both of which make allele fixation difficult. Even at the diploid level, inbreeding depression can limit genetic gains by reducing vigour and fertility. While homozygosity is important for trait stabilization and predictability, complete homozygosity is often unfeasible

due to the fitness costs of inbreeding. Interestingly, several studies have shown that maintaining some heterozygosity can actually improve traits such as yield, tuber number, and reproductive performance (Marand et al. 2019; Peterson et al. 2016; Phumichai and Hosaka 2006). This creates a trade-off in TPS breeding programs, which must balance the need for homozygosity with the risk of inbreeding depression. Fixation-Restitution Breeding offers a way to navigate this trade-off by enabling allele fixation at selected loci while preserving heterozygosity elsewhere, unlike traditional F1 hybrid models that rely on fully homozygous inbred lines.

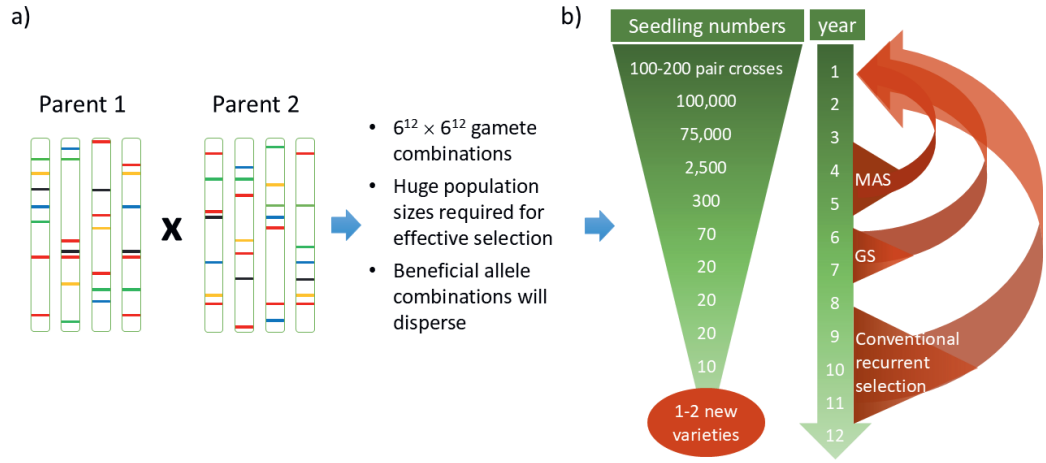


Figure 1. a) A chromosome diagram illustrating the meiotic outcomes of a cross between two highly heterozygous tetraploid parents. Each parent contributes four homologous chromosomes, with coloured segments representing different alleles or loci. Due to tetrasomic inheritance, meiosis in tetraploids produces an enormous number of potential gamete combinations. This extreme genetic variability necessitates large population sizes for effective selection, as recombination frequently disrupts favourable allele combinations, complicating the fixation of desirable traits. b) A schematic overview of the Teagasc/IPM Potato Breeding Programme, showing the progression from initial pair-crosses to advanced multi-location trials over an approximately 12-year cycle. The green triangle illustrates how the number of cultivar candidates narrows as selection advances. In addition to this conventional pipeline, the programme includes a Rapid Cycle scheme, in which progeny with stacked resistance (R) genes are identified early and recycled as parents to accelerate the accumulation of favourable traits. This integration of early-generation selection and rapid cycling shortens the breeding timeline while increasing genetic gain (adapted form Griffin et al. 2022).

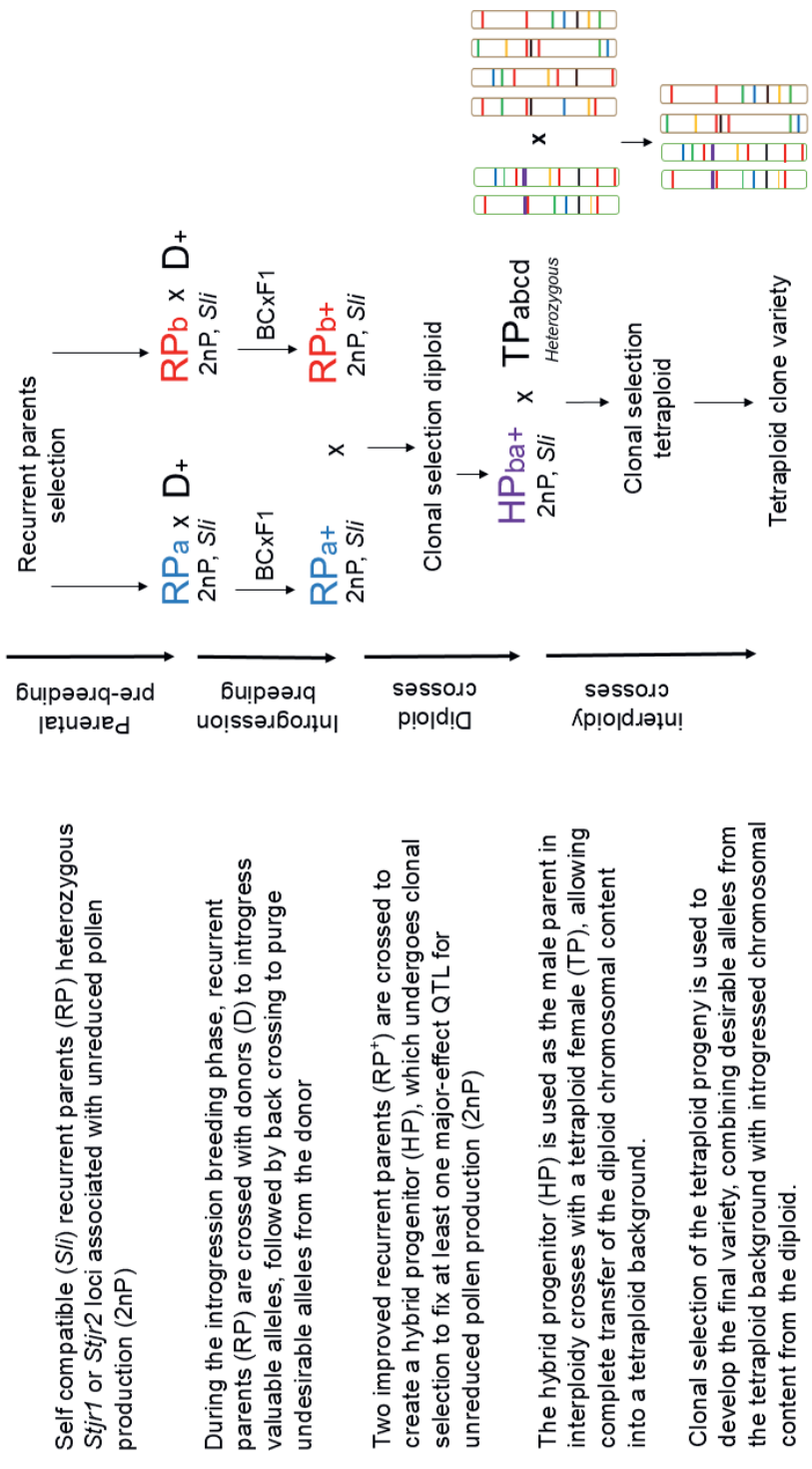


Figure 2. Overview of the Fixation-Restitution (Fix-Res) breeding strategy in potato, adapted from Clot (2023). The figure outlines key phases of the breeding scheme, including parental pre-breeding, introgression, diploid and interploidy crosses, and clonal selection. In this strategy, the fixation and introgression of key traits are performed at the diploid level to overcome the complexities of tetraploid inheritance, and the assembled genetic content is then transferred to the tetraploid level via chromosomal restitution. The shift toward genomic-assisted breeding.

The increasing complexity of modern breeding objectives, such as disease resistance, climate adaptation, tuber quality, and market-specific demands, has prompted the need for more predictive and efficient breeding strategies. One such strategy is the Breeding by Design approach, proposed by Peleman and Rouppe van der Voort (2003), which envisions a system in which favourable alleles across the genome are identified, tracked, and systematically combined into elite genotypes. The Breeding by Design concept provided a strategic framework for integrating genomic tools—such as molecular markers and association mapping, into breeding pipelines, building on decades of earlier marker development.

Markers in potato breeding

At the core of this framework is the use of molecular markers to accelerate genetic gain and improve selection accuracy. These tools enable breeders to make informed decisions early in the breeding cycle, before traits can be reliably evaluated in the field—when population sizes are large and seed tubers are limited.

Modern breeding programs increasingly incorporate genomic tools such as Genome-Wide Association Studies (GWAS), Marker-Assisted Selection (MAS) and Genomic Selection (GS), as part of design-based strategies. While several methods exist to identify trait-associated loci, including classical QTL mapping in biparental populations, GWAS has become particularly useful for exploring marker–trait associations in genetically diverse panels. These studies detect quantitative trait loci (QTL) and their linked markers, which serve as candidates for MAS. In MAS, breeders use validated markers to screen seedlings for specific favourable alleles using tools such as KASP or PCR-based Single Nucleotide Polymorphism (SNP) genotyping.

MAS is particularly effective when targeting major-effect QTL, for example, disease resistance (Gebhardt et al. 1993) or tuber shape (van Eck et al. 1994a). However, it becomes less useful for polygenic traits, where many small-effect loci contribute to overall performance. This is where Genomic Selection (GS) becomes essential. It is a form of MAS in which genetic markers spread across the whole genome are used to predict breeding values of individuals for the purpose of ranking selection candidates (Goddard and Hayes 2007; Heffner et al. 2009; Meuwissen et al. 2001). Unlike GWAS, which identifies specific QTLs, GS captures the combined effects of many loci, making it particularly effective for polygenic traits influenced by multiple small-effect genes. By incorporating genome-wide markers in the Genomic Prediction (GP) model, to estimate Genomic Estimated Breeding Values (GEBVs). These predictions allow breeders to rank and select individuals based on their expected performance, even before field trials begin (Slater et al. 2014). This approach is better suited for complex and low-heritability traits (Goddard and Hayes 2007; Heffner et al. 2009).

The key features and applications of GWAS, MAS, and GS are summarized in Table 1. Incorporating genomic tools like MAS and GS into breeding programs enhances selection

efficiency, shortens the breeding cycle, and allows for better resource allocation by focusing efforts on the most promising candidates earlier in the process (Bradshaw 2017). This kind of targeted selection is already being applied in breeding programs like Teagasc-IPM, where a Rapid Cycle scheme leverages MAS to identify and recycle progeny carrying multiple resistance genes early in the breeding process (Griffin et al. 2022; Figure 1b).

Table 1. Summary of key genomic tools used in potato breeding programs. Genome-Wide Association Studies (GWAS) identifies marker–trait associations to uncover the genetic basis of target traits. Marker-Assisted Selection (MAS) uses a limited number of validated markers to select individuals carrying favourable alleles. Genomic Selection (GS) applies genome-wide markers to predict breeding values for complex traits, enabling early and efficient selection.

	GWAS	MAS	GS
Purpose	Discover marker-trait associations	Select individuals based on key markers	Predict breeding values using genome-wide markers
Markers Used	Genome-wide markers	Few markers/QTL with major effects	All markers, including small-effect markers
Application	Identifies new QTL & candidate markers	Fast selection of high-value individuals	Predicts complex trait performance
Best For	Research & pre-breeding	Simple traits (e.g., disease resistance, tuber shape)	Polygenic traits (e.g., yield, chipping colour)
Breeding Impact	Provides genetic knowledge	Selects beneficial alleles in elite lines	Accelerates genetic gain across breeding cycles

GWAS have been increasingly applied in potato, particularly at the tetraploid level, to identify loci associated with agronomic traits in breeding-relevant material (Baldwin et al. 2011; Byrne et al. 2020; D'hoop et al. 2014; D'hoop et al. 2008; Klaassen et al. 2019; Li et al. 2010; Lindhout et al. 2011; Malosetti et al. 2007; Prodhomme et al. 2020; Rosyara et al. 2016; Schönhals et al. 2017; Sharma et al. 2018; Urbany et al. 2011; Vos et al. 2022; Zhang et al. 2022a). In contrast, most genetic studies in diploid potato have focused on mapping specific traits using biparental populations, with relatively few markers and smaller population sizes (Díaz et al. 2021; Parra-Galindo et al. 2021; Yang et al. 2021). As a result, GWAS studies in diploid germplasm sets remain limited in both scale and number.

The potential of GS to address the complexities of potato breeding has generated growing interest within the breeding community. A number of studies have assessed the performance of various statistical and machine learning methods for predicting GEBVs across a wide range of traits (Aalborg and Nielsen 2024; Aalborg et al. 2024; Byrne et al. 2020; Ortiz et al. 2022; Ortiz et al. 2023; Pandey et al. 2023; Selga et al. 2021a; Slater et al. 2016; Stich and Van Inghelandt 2018; Sverrisdóttir et al. 2017; Sverrisdóttir et al. 2018; Wilson et al. 2021). By enabling the prediction of untested genotypes based on genomic

profiles, GS improves selection accuracy and accelerates the development of superior potato varieties.

Evolution of marker systems

The introduction of molecular markers revolutionized plant breeding by enabling genetic diversity analysis, trait mapping, and selection efficiency. In potato, early marker systems included Restriction Fragment Length Polymorphism (RFLP) for genetic mapping (Bonierbale et al. 1988b; Gebhardt et al. 1993; Gebhardt et al. 1991; van Eck et al. 1994b), Simple Sequence Repeat (SSR) markers, also known as microsatellites, for high-resolution diversity studies (Milbourne et al. 1998; D'hoop et al. 2014; Ghislain et al. 2004), and Random Amplified Polymorphic DNA (RAPD) for genome-wide fragment analysis (Baird et al. 1992; Demeke et al. 1996). Amplified Fragment Length Polymorphism (AFLP) improved upon RAPD by enhancing resolution and reproducibility (van Os et al. 2006; Vos et al. 1995), while High-Resolution Melting (HRM) analysis later emerged as a cost-effective genotyping method (Villano et al. 2015). Despite their contributions and their role in laying the foundation for modern genomics, these methods were limited by low genome coverage, poor scalability, and high costs, reducing their feasibility for large-scale breeding programs.

Kompetitive Allele Specific PCR (KASP) is a simplified fluorescence-based methodology to genotype specific polymorphisms or INDELS (https://excellenceinbreeding.org/) (Lindhout et al. 2011), and is widely adopted as an effective diagnostic tool in potato breeding (Sood et al. 2022; Meade et al. 2020b; Prodhomme et al. 2020; Clot et al. 2020; Sorensen et al. 2023; Asano and Endelman 2024). However, while KASP assays offer a cost-effective solution for targeted marker genotyping, they lack genome-wide coverage and are not optimal for broader applications in genomic selection.

The post-genome era has seen the rapid expansion of high-throughput genotyping technologies, enabling breeders to analyse thousands of markers simultaneously. SNP arrays have become a widely used tool in genomic-assisted breeding, providing high-density genome-wide marker coverage. SNPs are co-dominant markers that are abundant, widespread across the genome, and exhibit a high degree of polymorphism (Uitdewilligen et al. 2013). In potato, the Infinium SNP array has been the main genotyping tool for potato genetics, starting with 8,303 markers and expanding to a 12K version (Hamilton et al. 2011; Felcher et al. 2012). Later versions included the 22K V3 array, incorporating SNPs from 83 tetraploid varieties (Uitdewilligen et al. 2013; Vos et al. 2015), and the 31K V4 array, further increasing marker coverage (Sharma and Bryan 2017). While SNP arrays offer high-throughput genotyping, they require prior sequence information, have a high setup cost, and demand advanced technical expertise. Additionally, the reliance on predefined SNPs can introduce ascertainment bias (Vos et al. 2015). SNP arrays are limited to **bi-allelic variants**, which can reduce the ability to fully capture the genetic diversity present in highly heterozygous species like potato. SNP array-based platforms are only cost-effective for large sample sizes, making them primarily accessible to large commercial breeding programs or through sample pooling within international consortia.

Genotyping-by-Sequencing (GBS) emerged as a flexible and affordable alternative, allowing simultaneous SNP discovery and genotyping. It reduces genome complexity using restriction enzymes and is compatible with diverse germplasm without prior marker development (Elshire et al. 2011). GBS has been successfully applied in multiple potato studies for trait mapping and genomic prediction (Byrne et al. 2020; Sverrisdóttir et al. 2017; Wilson et al. 2021; Wang et al. 2021; Bastien et al. 2018; D'hoop et al. 2008; Kloosterman et al. 2013; Lindqvist-Kreuze et al. 2014; Rosyara et al. 2016; Schönhals et al. 2016; Sharma et al. 2018; van Eck et al. 2017). However, GBS with restriction enzymes is covered by a patent (http://www.google.com/patents/US8815512) and the cost of GBS, when commercially sourced, ranges from 50 to 100 euros per sample, depending on sample numbers and coverage (Byrne et al., unpublished data). GBS with restriction enzymes is also prone to low read-depth and missing data complicating downstream bioinformatics.

While SNP arrays, GBS, and KASP have significantly advanced genomic-assisted breeding, they also present limitations in cost, scalability, and marker resolution. Large-scale breeding programs often require thousands of samples to be genotyped annually, making affordability and high-throughput efficiency crucial factors in selecting a genotyping platform. Additionally, SNP arrays are often affected by ascertainment bias due to their required development phase (Vos et al. 2015). To address these challenges, multiplexed amplicon-based genotyping approaches have been developed to provide cost-effective and scalable alternatives, allowing for targeted SNP analysis without requiring whole-genome sequencing.

Toward affordable and scalable genotyping solutions

Recently, several groups have developed targeted genotyping-by-sequencing (GBS) approaches based on multiplex amplicon sequencing. These methods allow for the simultaneous amplification and sequencing of multiple specific genomic regions in a PCR reaction, significantly enhancing efficiency and cost-effectiveness. They enable the detection of multiple SNPs or other genetic variations across the genome without requiring whole-genome sequencing.

One such technology is DArTag, a targeted genotyping platform developed by Diversity Arrays Technology (DArT) (Jaccoud et al. 2001). DArTag employs molecular inversion probes (MIPs) to provide cost-effective, high-throughput SNP analysis across various polyploid crops, including strawberry (Hardigan et al. 2023), alfalfa (Medina et al. 2025), blueberry (Zhao et al. 2024b), sweet potato (Zhao et al. 2024a), and potato (Endelman et al. 2024). This platform facilitates genomic studies and breeding programs by offering mid-density marker panels tailored to the genomic characteristics of each species. In potato, DArTag markers were selected from the potato Infinium array, with the V1 version including 2.5K markers and the V2 expanding to 4K markers covering additional resistance and trait loci. With a cost of 12–17 USD per sample (depending on multiplexing levels, (https://excellenceinbreeding.org/). A similar system, Solseq amplicon sequencing, has been developed at Wageningen University. This method, based on 2,880 amplicons of pre-

defined SNP array, also offers a potential cost-effective tool for genomic-assisted breeding at a cost of 12 euros per sample excluding DNA isolation and data processing (personal communication dr. H.J. van Eck, Plant Breeding, WUR).

Another commercially available targeted GBS platform is Flex-Seq, developed by LGC (https://www.biosearchtech.com/flex-seq). Flex-Seq utilizes a multiplex PCR approach to selectively amplify a predefined set of SNP loci, ensuring high reproducibility and accurate allele dosage estimation. The process involves two steps: an initial PCR to amplify target genomic regions, followed by a second PCR that adds sequencing adapters and sample barcodes for multiplexing. LGC Biosearch Technologies offers a pre-designed 22K Flex-Seq loci panel specifically for potatoes.

Although DArTag, Flex-Seq, and Solseq provide cost-effective amplicon-based genotyping solutions, they are still dependent on pre-selected SNP arrays, expanding the SNP target set in these platforms requires additional research, development, and associated costs, which may limit their scalability for broader breeding applications.

Capture-seq is a hybridization-based targeted sequencing approach that enriches specific genomic regions before sequencing, improving genotyping efficiency and variant detection. Unlike SNP array-based approaches, Capture-Seq flexibly targets SNPs, genes, and QTLs and is adaptable, making it valuable for breeding activities. It has been used in a GWAS study in potato (Angelin-Bonnet et al. 2023). Panels are designed and it is commercially available for other crops but not yet for potato (https://www.biosearchtech.com/capture-seq).

In wheat, MRASeq (Multiplex Restriction Amplicon Sequencing) was developed as a high-throughput genotyping method that reduces genome complexity by selectively amplifying genomic regions flanked by restriction enzyme sites. This approach utilizes a two-step PCR-based protocol: the first step involves restriction enzyme digestion and targeted amplification, followed by a second PCR that adds sequencing primers and sample barcodes for multiplexing. Unlike SNP arrays, MRASeq is an NGS-based genotyping method that identifies SNPs and genetic markers through targeted PCR amplification of genomic regions flanked by restriction sites. The cost of MRASeq is approximately 8 euros per sample (excluding bioinformatics and labour costs) (Bernardo et al. 2020).

Table 2. Comparison of molecular marker technologies for potato breeding. This table provides an overview of various molecular marker technologies used in potato breeding, highlighting their principles, marker density, cost, technical complexity, coverage, scalability, and availability.

Method	*Principle & Technology*	Marker Density	Cost per Sample	Complexity & Bioinformatics	Flexibility (Customization)	Scalability	Advantages	Limitations	Available Potato Panels	Offered as commercial service
SNP Arrays	***Targeted*** *hybridization-based detection of predefined SNPs (fluorescent readout)*	**Medium to high** (1K-100K SNPs)	**High** ($30-$300, depending on marker and sample size))	**High** (specialized equipment, standardized bioinformatic pipeline)	**Low** - Fixed SNP panel, limited customization; predefined marker selection	**Low**	High-throughput, reproducible	Expensive, fixed markers, Ascertainment Bias	Potato Infinium 8K (Hamilton et al. 2011), 12K (Felcher et al. 2012), 22K V3 (Uitdewilligen et al. 2013; Vos et al. 2015), 31K V4 (Sharma and Bryan 2017) arrays	Yes – Various providers
GBS (Genotyping by Sequencing)	*NGS-based genome reduction using* ***random, genome-wide****, restriction enzymes (NGS readout)*	**Very High** (>100K SNPs)	**Medium** ($30-$100)	**Medium** (NGS & custom bioinformatics and HPC)	**High** - Customizable sequencing targets; allows new SNP discovery	**Medium**	Cost-effective, genome-wide coverage, suitable for large populations	High missing data, requires imputation, expensive for commercial use		Yes
KASP Assays	***Targeted*** *Allele-specific PCR with fluorescence detection*	**Low** (1-100 markers)	**Low** ($1-$5)	**Low** (Simple PCR setup, and Direct allele calling via fluorescence)	**Low** - Fixed marker selection; limited to known SNPs, but easy to apply	**High**	Inexpensive, rapid and simple workflow, scalable for MAS	Limited to known markers. Not applicbple for GS		Yes – Various providers
DArTag (Diversity Array Technology)	***Targeted*** *multiplex amplicon sequencing, based on 2 PCR steps for probe hybridisation (NGS Illumina readout)*	**Medium** (1K-5K loci/SNPs)	**Low** ($17)	**Medium** (NGS & bioinformatics pipeline for amplicon sequencing)	**Medium** - Some customization possible; predefined but adaptable SNP panels	**High**	Good balance of cost and marker density, cost effective for breeding purposes	Requires design and predefined SNP panel	Potato DArTag V1 (2.5K SNPs), V2 (4K SNPs) (Endelman et al. 2024) (www.diversityarrays.com)	Yes – Diversity Arrays Technology (DArT)

Method	*Principle & Technology*	Marker Density	Cost per Sample	Complexity & Bioinformatics	Flexibility (Customization)	Scalability	Advantages	Limitations	Available Potato Panels	Offered as commercial service
Solseq	***Targeted** amplicon sequencing via Ion Torrent, targeting predefined SNP*	**Medium** (2.8K loci/SNPs)	**Low** ($12)*		**High** - Highly adaptable, with predefined panels that can be adapted with new markers.	**High**	Cost-effective for breeding purposes	Requires design and predefined SNP panel	Solseq potato panel (2,880 amplicons)	Developed at Wageningen University, not widely commercialized
Flex-Seq	***Targeted** GBS-based sequencing, multiplex amplicon sequencing targeting predefined **genome-wide** SNP loci. based on 2 PCR steps for probe hybridisation (NGS Illumina readout)*	**Medium to high** (2K-30K SNPs)	not publicly available	**Medium** (NGS & bioinformatics pipeline for amplicon sequencing)	**High** - Highly adaptable, with predefined panels that can be adapted with new markers.	**High**	Flexible, adaptable for different applications. High accurcy. - cost effective for breeding purposes.	Requires design and predefined SNP panel	22K Flex-Seq loci panel for potatoes (www.biosearchtech.com)	Yes – LGC Biosearch Technologies
Capture-seq	***Targeted** GBS-based sequencing, multiplex amplicon sequencing targeting **genome-wide** exons. based on 2 PCR steps for probe hybridisation (NGS Illumina readout)*	**Very High** (5K-200K SNPs)	not publicly available	**Very High** (advanced sequencing & alignment)	**High** - Genome-wide, highly adaptable; can be fine-tuned for different applications	**High**	High accurcy. Minimal ascertainment bias, compatible with complex genomes	Complex design and bioinformatics	NO	LGC Biosearch Technologies
MRASeq (Multiplex Restriction Amplicon Sequencing)	*Multiplex restriction amplicon sequencing, based on 2 PCR steps selectively amplifying genome regions flanked by **random**, **genome-wide** restriction sites. Ion Torrent sequencing.*	**Medium** (1K-10K SNPs)	low ($8)**	**High** (NGS & specialized setup, bioinformatics pipeline for amplicon sequencing)	**High** - Designed for flexible target capture; allows multiplexing and sample pooling	**High**	Low-cost, uniform genome coverage compared to GBS	Requires specialized library preparation	NO	No, Developed in research settings in wheat

**Excluding DNA isolation and data processing. **excluding bioinformatics and labour costs.*

Beyond bi-allelic SNPs: The potential of multi-allelic haplotypes

Genetic improvement in potato has traditionally relied on SNP-based analysis. In potato, large portions of the genome exhibit high heterozygosity and a high SNP density (Uitdewilligen et al. 2013). However, because SNPs are primarily bi-allelic, they often fail to capture the full complexity of allelic variation, particularly in the highly heterozygous and polyploid genome of crops like potato, where multiple alleles at a single locus interact to influence phenotypic traits. Haplotypes, defined as contiguous sets of genetic variants (SNPs, indels, or structural variants) inherited together from a single parent, offer a more comprehensive view of genetic variation (Figure 3). Since recombination does not occur randomly across the genome, haplotypes provide more meaningful genetic information than individual SNPs, better reflecting the true inheritance of genetic variation (Meuwissen et al. 2014). While SNPs are typically selected based on moderate to high minor allele frequency (MAF), meaning they primarily represent older mutations, new mutations that are often present at low frequencies may be lost before reaching detectable levels in a population (Meuwissen et al. 2014). Haplotypes offer greater discriminatory power in quantitative genetics because they are assumed to be in stronger linkage disequilibrium (LD) with QTL than individual SNP alleles (Calus et al. 2008; Hess et al. 2017; Meuwissen et al. 2014). As a result, haplotypes have the potential to serve as more effective markers for MAS and GS in genetic studies and breeding applications.

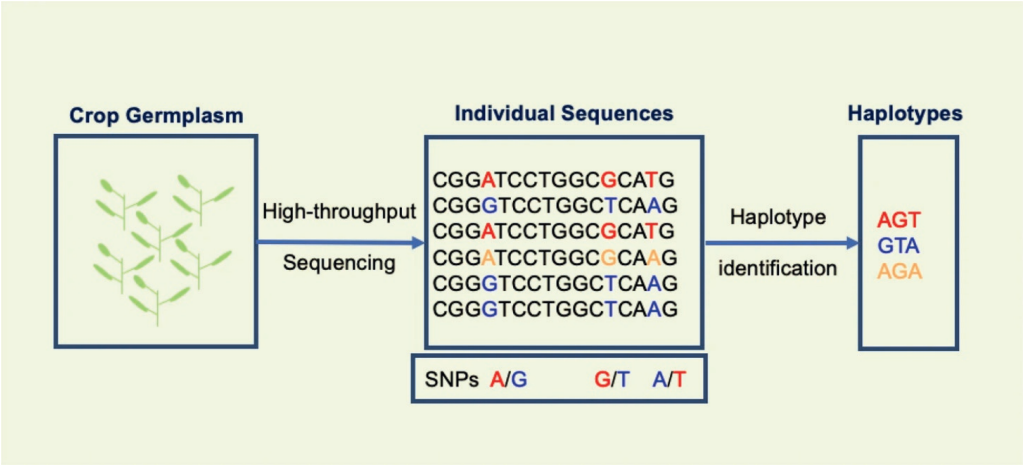


Figure 3. A demonstration of how resequencing crop germplasm enables the identification of polymorphic SNPs, which are subsequently used to develop multi-allelic haplotypes (Bhat et al. 2021)

Haplotypes can be categorized as long or short, each offering distinct advantages in genetic studies. In animal breeding, haplotypes are typically constructed using large SNP arrays and inferred statistically by phasing SNPs based on known parental inheritance. These methods can result in long haplotypes, sometimes spanning entire chromosomes, and are implemented with software such as BEAGLE (Browning and Browning 2009). Long haplotypes tend to remain intact over multiple generations due to selection pressures that

preserve beneficial allele combinations. Some studies suggest that breeders retain longer haplotype blocks than expected by chance (Fradgley et al. 2019), likely due to selection against recombinants that break apart co-adapted or beneficial allele combinations (Brinton et al. 2020). These long-range haplotypes, often referred to as "founder haplotypes" (Scott et al. 2021), can be useful for tracking recombination events and identifying ancestral genetic contributions in breeding programs. However, long haplotypes tend to be rarer in a population, which can reduce statistical power in association studies.

Short haplotypes, in contrast, represent smaller genomic blocks that are more frequently reshuffled by recombination (Brinton et al. 2020). These short multi-allelic genomic variants are often more informative for fine-mapping studies, as they provide higher resolution when identifying causal variants linked to traits of interest (Difabachew et al. 2023; Jiang et al. 2018; Lu et al. 2011; Sallam et al. 2020). While research on haplotype applications in plant breeding remains limited, several studies have demonstrated their potential as multi-allelic markers in GWAS and GS across various crops, including canola, maize, wheat, soybean, ryegrass (Difabachew et al. 2023; Jiang et al. 2018; Lu et al. 2011; Sallam et al. 2020; Ledesma-Ramírez et al. 2019; Lu et al. 2012; Sehgal et al. 2020; Weber et al. 2023; Kang 2023). In potato, haplotypes have been used to identify genomic regions associated with traits such as disease resistance (van Eck et al. 2017) and protein content (Kloosterman et al. 2013). Thérèse Navarro et al. (2022) developed a software, mpQTL, for QTL analysis at any ploidy level under bi-allelic and multi-allelic models in multiparental populations. However, the broader application of multi-allelic haplotypes in genetic studies and practical potato breeding remains largely unexplored. While haplotype-based approaches, including their use in GWAS and genomic selection, have been proposed, their effectiveness in breeding programs has yet to be fully tested.

Haplotype construction can be achieved through either statistical phasing or read-backed haplotyping. Statistical phasing, which involves ordering SNP markers based on consensus map positions (Jiang et al. 2018; Sallam et al. 2020) or using LD-based statistical methods (Difabachew et al. 2023) with tools such as BEAGLE (Browning and Browning 2009) and Haploview (Barrett et al. 2005). However, these approaches face challenges, particularly when marker density is low or in low-coverage sequencing scenarios (Schaumont et al. 2022). Phasing multi-allelic haplotypes in tetraploid, highly heterozygous potato is especially challenging due to the presence of four homologous chromosome copies, leading to allele dosage ambiguity, complex recombination patterns, and higher phasing error rates compared to diploids, where bi-allelic variants represent only two haplotype alleles of the genome. In contrast, read-backed haplotyping, constructs haplotypes directly from sequencing reads rather than relying on computational inference or predefined SNPs. This approach provides a more direct and reliable way to infer haplotypes from sequencing data, reducing errors associated with statistical phasing (Schaumont et al., 2022).

Scope and outline of the thesis

This work was conducted as part of the DIFFUGAT project (https://diffugat.eu/), which focuses on the development of Fix-Res breeding. A key aspect of Fix-Res breeding is the requirement for repeated rounds of self-fertilization and selection to develop recurrent parents that have undergone purging of lethal and excessively deleterious alleles. Additionally, during the deployment phase of Fix-Res breeding, efficient selection for numerous beneficial alleles related to disease resistance, abiotic stress tolerance, and tuber quality is required. Following introgression of these traits, recurrent backcrossing and selection further refine the genetic composition of elite lines. Thus, there is a strong need for an affordable, high-resolution genotyping system that balances costs, accuracy, and scalability for modern potato breeding applications, and particularly in development of recurrent parents for fix-res breeding.

This thesis focuses on the development, application and evaluation of a cost-effective genotyping platform, that leverages haplotype-based information while minimizing ascertainment bias and increasing marker flexibility to enhance genetic studies for different traits including quantitative ones and improve breeding efficiency in potato. The research explores the utility of haplotype-based marker analysis in key areas such as genome-wide association studies (GWAS), inbreeding tracking, and genomic prediction (GP) to support modern breeding strategies, particularly in the context of Fix-Res breeding.

In Chapter 2, we introduce PotatoMASH (Potato Multi-Allele Scanning Haplotags), a novel low-cost genome-scanning marker platform designed for cost-efficient, high-throughput genotyping. We developed a panel of 339 multi-allelic regions spaced at 1 Mb intervals throughout the euchromatic portion of the genome, genotyped using a multiplex amplicon sequencing approach followed by deep NGS sequencing (2x150bp Illumina sequencing). This system enables large-scale genotyping at a cost of 5 EUR per sample (excluding labour and overhead costs). In conjunction with Stack Mapping Anchor Points (SMAP) software for read-backed haplotyping, PotatoMASH generates two types of markers, SNPs and short-read multi-allelic haplotypes (haplotags) based on these SNPs. This chapter describes the development and optimization of the platform, detailing marker selection, sequencing protocols, and bioinformatics processing to generate the two marker sets. We applied PotatoMASH to a population of over 700 potato lines, generating 2279 SNPs and 2012 haplotags across 334 loci, with haplotype diversity ranging from 2 to 14 haplotypes per locus. The platform was successfully used for diverse applications, including diagnosing pest-resistance markers, constructing genetic maps, and tracking genetic variation in a diploid segregating population. Furthermore, a GWAS for fry colour in a tetraploid potato population using SNPs and SNP-based haplotags revealed that, while the PotatoMASH SNP set failed to detect any QTL, the haplotag set successfully identified the same QTL as a previous GWAS using 43.6K GBS-derived SNP markers. This result suggests that haplotags may offer superior discriminatory power for QTL detection in genome-wide association studies.

In Chapter 3, we used PotatoMASH to genotype a large panel of diploid clones that served as the foundation of Fix-Res breeding. This diploid panel was phenotyped for 23 agronomic and quality traits over three years as part of a large scale collaborative effort across breeding programmes. We performed GWAS to identify QTL for all traits, utilizing both SNPs and short-read haplotypes (haplotags) based on combinations of those SNPs, derived from read-backed phasing. This enabled us to identify a total of 37 unique QTL across both marker types. Although we initially hypothesized that haplotags may offer better discriminatory power than SNPs for QTL detection in GWAS, interestingly, the haplotags did not consistently outperform bi-allelic SNPs in QTL detection for all traits. A core of 10 QTL were detected with both SNPs and haplotags. Haplotags enabled the detection of 14 additional QTL not found using SNPs, while the bi-allelic SNP set identified 13 QTL that were not detectable with haplotags. These findings suggest that both marker types should be routinely used in parallel to maximize QTL detection power. We report 19 novel QTL for nine traits: Skin Smoothness, Sprout Dormancy, Total Tuber Number, Tuber Length, Yield, Chipping Colour, After-cooking Blackening, Cooking Type and Eye depth.

In Chapter 4, we used PotatoMASH for tracking homozygosity in diploid breeding. Utilising a collection of 271 inbred diploid clones from the Wageningen diploid breeding program, we obtain a "snapshot" of the genetic composition shaped by over 40 years of breeding efforts. Furthermore, we examine a self-compatible individual lineage of three generations of inbreeding from the program to identify key hotspots of heterozygosity across chromosomes. We assess the effectiveness of PotatoMASH-derived haplotags in evaluating genome-wide homozygosity changes and inferring selfing rates in the breeding materials by comparing the resolution of haplotags versus SNPs. A major focus of this study is the identification and characterization of residual heterozygosity (RH) regions, which provides crucial insights into selection pressure and genetic stability in breeding programs. Assessing the patterns of allele fixation and heterozygosity across three generations of selfing, we observed average homozygosity levels of 82-83.4% for Identity-by-State (IBS) and 72.6-74.8% for Identity-by-Descent (IBD) in S3 progenies, which were lower than anticipated. Hotspots of heterozygosity were detected across all chromosomes, with chromosome 5 remains entirely heterozygous after three generation of selfing and chromosome 11 reaching full homozygosity in one S3 progeny. We discuss how luck of allele fixation on chromosome 5 may be associated with reproductive-related QTL and genes that favour the heterozygous state due to potentially deleterious alleles.

In Chapter 5, we assess the effectiveness of haplotags in genomic prediction (GP). This chapter evaluates the predictive accuracy (PA) of haplotags in genomic selection models, comparing their performance with SNP-based approach. In a tetraploid population, we compared the PA of PotatoMASH data with that of GBS data, specifically for the complex trait fry colour. The results showed that by using only the markers generated by PotatoMASH (2,236 SNPs compared to 43.6k SNPs generated by GBS), the PA was moderately reduced by 14% when using SNPs, and only 9% by using the 2,000-3,390 haplotags, making PotatoMASH a cost-effective solution for large-scale breeding programs. We also tested

PotatoMASH performance across 23 agronomic, quality, and morphological traits in a diploid panel, with PA ranging from medium to high (0.29-0.81). Haplotags enhanced PA for 11 traits compared to SNPs of the same data. In contrast, GP based on individual SNPs performed better for other six traits. In this chapter, we also explored an alternative haplotyping approach that can potentially capture more genetic diversity, using SMAP software, we compared two haplotype-calling approaches: One based on variants at pre-called SNP locations on the haplotag (*haplotype-sites*) that we have used in the previous chapters and a new module of direct read-mapping method (*haplotype-window*). PA with the haplotags derived from *haplotype-window* outperform in 6 traits compared to the haplotags derived from haplotype-site.

In Chapter 6, I reflect on the main findings of this thesis and discuss how PotatoMASH can support key decisions within the Fix-Res framework and in broader efforts toward Breeding by Design in potato. I compare its performance to other genotyping platforms, examine its current limitations, and outline practical next steps to improve its accessibility and impact in routine breeding applications.

Chapter 2

PotatoMASH - A Low Cost, Genome-Scanning Marker System for Use in Potato Genomics and Genetics Applications

Authors

Maria de la O. Leyva-Pérez[1], Lea Vexler[1,2], Stephen Byrne[1], Corentin R. Clot[2], Fergus Meade[1], Denis Griffin[1], Tom Ruttink [3], Jie Kang [1,4], Dan Milbourne[1]

Affiliations

[1]Teagasc, Crop Science Department, Oak Park, R93 XE12 Carlow, Ireland

[2]Plant Breeding, Wageningen University & Research, P.O. Box 386, 6700 AJ Wageningen, The Netherlands

[3]Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Plant Sciences Unit, Caritasstraat 39, Melle 9090, Belgium

[4]Beef + Lamb New Zealand Genetics, 3 Crawford Street, PO Box 5501, Dunedin, 9054, New Zealand

Published in Agronomy (2022)

https://doi.org/10.3390/agronomy12102461

Abstract

We have developed PotatoMASH (Potato Multi-Allele Scanning Haplotags), a novel low-cost, genome-scanning marker platform. We designed a panel of 339 multi-allelic regions placed at 1 Mb intervals throughout the euchromatic portion of the genome. These regions were assayed using a multiplex amplicon sequencing approach, which allows for genotyping hundreds of plants at a cost of 5 EUR/sample. We applied PotatoMASH to a population of over 700 potato lines. We obtained tetraploid dosage calls for 2012 short multi-allelic haplotypes in 334 loci, which ranged from 2 to 14 different haplotypes per locus. The system was able to diagnose the presence of targeted pest-resistance markers, to detect quantitative trait loci (QTLs) by genome-wide association studies (GWAS) in a tetraploid population, and to track variation in a diploid segregating population. PotatoMASH efficiently surveys genetic variation throughout the potato genome and can be implemented as a single low-cost genotyping platform that will allow the routine and simultaneous application of marker-assisted selection (MAS) and other genotyping applications in commercial potato breeding programmes.

Introduction

Numerous applications in plant genetics, genomics, and breeding are based on genome-wide marker analysis, and although genotyping costs have dropped considerably over the recent years, they are still one of the major barriers in applications requiring the generation of genome-wide marker data for large numbers of samples. Applications such as genome-wide association studies (GWAS), genomic selection (GS), and marker-assisted selection (MAS) routinely involve sample sets in the thousands, and breeding applications have the additional cost burden of iterative application to potentially thousands of genotypes over generations. Even smaller-scale applications, such as genetic mapping in bi-parental populations at low-to-medium resolution, for gene and QTL discovery purposes benefit from a lower cost base, and there are numerous advantages in terms of the cross-study comparability of utilising the same platform for higher- and lower-throughput experiments.

We decided to explore the potential for developing a single, highly economical, yet reasonably powerful approach to genome-wide genotyping that could be applied broadly to all of the above applications, adopting a design-criterion-guided approach to create such a platform. The goal of this study was to utilise the information and criteria described below to design an extremely cost-effective, genome-wide genotyping system in potato. At the outset, we set a cost per assay benchmark of approximately EUR 5 per sample (excluding labour costs), with the goal that the assay should be technically feasible to carry out in a standardly equipped molecular biology laboratory setting. The process should be applicable to a dynamic range of samples from hundreds to the low-thousands with a low requirement for automation. The assay should have a core set of loci that enable the “scanning” of genetic variation across the genome at a density that is likely to be able to detect variants underlying phenotypic characteristics measured in a population. In addition, it should be expandable, and specific loci of interest to the user should be easily added to the platform.

Numerous genotyping systems have been applied to potato: all have advantageous design features for different situations, but none possess all of the above features. “Predesigned” systems such as arrays have the advantage of offering the user easy access to a community-wide SNP set, including relatively easy data capture pipelines. Problems with ascertainment bias experienced in arrays can be addressed by utilising a sufficiently broad germplasm set at the design phase (Hamilton et al. 2011; Vos et al. 2015). However, arrays require a separate SNP discovery phase and survey a fixed set of polymorphisms, making them less adaptable. Genotyping-by-sequencing (GBS) using restriction enzyme-based genome reduction approaches has also been applied in potato (Byrne et al. 2020; Sverrisdóttir et al. 2017). This approach does not require prior sequence information and is partially “tunable” in terms of the total number of loci covered. In contrast to the relatively simple and straightforward library preparation, GBS data analysis is complicated by the nature of the random location, its reduced-representation approach, it generates a large proportion of missing data, and it requires several statistical assumptions to be made in order to call variants (Wickland et al. 2017). Sequence capture-based GBS approaches

have also proven to be powerful in potato, both to survey variation on a genome-wide level (Uitdewilligen et al. 2013) or to target specific motifs such as resistance loci (Jupe et al. 2013). These approaches conform to some of the criteria above but tend to be technically onerous. Costs for all of these approaches exceed the criterion set above, with the exact price per assay varying on the basis of a large number of variables.

Recently, Campbell et al. (2015) demonstrated the utility of a low-cost, PCR-based, genome-wide approach called GT-Seq (Genotyping in Thousands by Sequencing) in trout. This approach seemed to have the potential to embed many of the criteria described above, one of the most attractive features being the low per-sample cost of USD 5 when library construction steps are performed by the user. Briefly, GT-Seq uses two thermal cycling steps for the multiplexed amplification of relatively small panels (50-500) of short loci of 100-200 bp containing targeted single-nucleotide polymorphisms (SNPs). During this process, sequencing adapters and dual-barcode sample-specific sequence tags are incorporated into the amplicons, enabling thousands of individuals to be pooled into a single library to be sequenced in an Illumina HiSeq Lane. We decided to use the GT-Seq approach as a platform to develop a low-cost genotyping system, using existing knowledge of the nucleotide diversity and LD structure in potato.

The density of genetic markers is an important feature in genome-wide marker systems. Whilst it seems technically feasible to include thousands of loci in the GT-Seq assay approach, we focused on minimising the number of loci surveyed to reduce cost and with a view to a greater technical achievability. This begs the question: what is the minimum number of loci that would provide reasonable genome coverage of potato taking into account that the most frequent application for molecular markers is gene discovery or tracking of allelic variants in populations? In potato, it has been found that “useful” levels of LD extend between 0.6 and 1.5 Mb depending on the population under examination and the LD criterion used (Sharma et al. 2018; Vos et al. 2017). Significantly, there is also almost no LD decay observed across the entire span of the pericentromeric heterochromatin, which accounts for approximately 50% of the genome in potato. Thus, “complete coverage” of the genome could theoretically be achieved by efficiently surveying variation at ~400 loci evenly distributed every 1 Mb across the euchromatic portion of the 840 Mb genome (Consortium 2011), so no site could be more than 0.5 Mb from at least one locus. However, SNPs are almost entirely bi-allelic, and surveying a single SNP locus per megabase will not efficiently survey the diversity of real haplotypes at any one locus. This problem is especially pronounced in potato, where large parts of the genome exhibit a high degree of heterozygosity. For instance, for the recent haplotype-resolved genome sequence of the diploid line RH89-039-16, the average SNP polymorphism rate between the two haplotype genomes was estimated at approximately 1 in 50 nucleotides for syntenic regions. When looking across multiple haplotypes, this rate can actually increase, and polymorphism rates of between 1/25 and 1/15 were observed for non-coding and coding regions, respectively, by Uitdewillegen et al. (2013). This high polymorphism rate is reflected by a high level of allelic or haplotypic diversity in potato germplasm. For example, using a targeted

resequencing approach, Uitdewillegen et al. (2013) were recently able to identify 16 allelic variants of the Glucan Water Di-kinase (GWD gene) by aggregating information from 81 SNPs over two regions, totalling 1 kb of the 16.5 kb. Using a variety of nucleotide windows generally under 1000 kb, it seems that gene haplotype numbers range between 5 and 20 in potato (Uitdewilligen 2012; Uitdewilligen et al. 2022; Uitdewilligen et al. 2013; Wolters et al. 2010). Thus, whilst in terms of LD structure, the concept of surveying polymorphism at 400 × 1 Mb intervals might make sense, the actual number of relatively evenly distributed bi-allelic SNPs at which variation is surveyed is likely to have to be at least 10-fold higher in order to capture the majority of the haplotypes present in any potato germplasm collection.

Interestingly, this high level of nucleotide diversity in potato also suggests a technical approach to minimising the number of loci to be analysed to achieve good coverage at an allelic diversity level. An interesting feature of resequencing data data is that the polymorphic content of individual reads, read pairs, or processed tags can be aggregated into what Tinker et al. (2016) referred to as “Tag-level haplotypes” or haplotags (Tinker et al. 2016). Haplotags may contain multiple SNPs, especially in an SNP-dense species such as potato, and these differing combinations of bi-allelic SNPs over the length of the tag or read produces an alternative set of genotypes that better reflect the real underlying allelic (or short-range haplotypic) variation at that locus. Tinker et al. (2016) utilised this concept for the software package Haplotag, which implements a reference-free approach for capturing this type of variation from resequencing data data and has subsequently been used in oats and other species (Baral et al. 2020; Canales et al. 2021; Tinker et al. 2016).

In this manuscript, we describe the development of the PotatoMASH (Potato Multi-Allele Scanning Haplotags) tool, a GT-Seq-based genotyping platform designed on the above principles. The goal of PotatoMASH is to converge low per-sample cost with reasonable genotyping power across multiple applications for potato breeding and genetics. This iteration of PotatoMASH is based on surveying SNP variation in NGS reads across 339 loci spread across the euchromatic portion of the potato genome at 1Mb intervals according to the DM reference pseudomolecule assembly (Consortium 2011). Because of the availability of a reference pseudochomosome molecule-scale assembly in potato, we utilised a novel algorithm called SMAP (Stack Mapping Anchor Points) (Schaumont et al. 2022), which is designed for stacked NGS reads, including those generated by highly multiplex amplicon sequencing approaches. In order to test the scalability and adaptability of the system, 10 loci containing diagnostic SNP loci for resistance to pests and pathogens were also included into the amplicon panel. We tested the ability of PotatoMASH combined with the SMAP haplotype calling pipeline to reveal short-range allelic diversity at the target loci in a tetraploid potato population, comprising 765 independent genotypes accumulated from the third field generation of a commercial potato breeding programme and in a diploid bi-parental mapping population comprising 92 F1 progeny individuals. In the tetraploid population, we demonstrated the apparent superior ability of ~2000 haplotag-derived allelic variants to detect a previously mapped QTL for fry colour (Byrne et al. 2020) relative to the component bi-allelic SNP variants used to derive these haplotags. In the diploid population, we

demonstrate the ability of PotatoMASH to generate a contiguous, haplotype-resolved genetic map of potato. Finally, we discuss the characteristics and potential future utility of PotatoMASH and similar approaches for potato breeding and genetics.

Materials and Methods

The PotatoMASH primer design process was carried out in 2018 when the DM_v4.04 was the latest version. For accuracy, we describe the process as performed using that version throughout the manuscript, facilitating comparisons to the study of Byrne et al. (2020), which also used that version. We provide the bed file used in this work with PotatoMASH loci coordinates for DM_v4.04 as Supplementary Materials (File S2). For utility with the current V6.1 genome, we include in the same file the loci coordinates according to DM_v6.1, which facilitates the future haplotype analysis with PotatoMASH in DM_v6.1.

PotatoMASH Primers Panel Design

First, we defined the euchromatic portion of the genome to be targeted (Table 1) and set the boundaries of euchromatin/heterochromatin based on previous knowledge of the genetic architecture of potato and recombination frequencies (Consortium 2011; Sharma et al. 2013; Tang et al. 2009).

We mapped WGS re-sequencing data of 75 commercial cultivars (Meade et al. 2020a); 33× coverage, 5 pools of 15 cultivars) to the *Solanum tuberosum* genome DM_v4.04 (Hardigan et al. 2016) using BWAMEM (Li 2013). We used Popoolation software (Kofler et al. 2011) to calculate the number of SNPs per 500 bp window (minimum coverage 20× and fraction of allele frequency 0.9). We selected regions of 10–30 SNPs/500 bp window to be explored with IGV software (Robinson et al. 2011) set to highlight variants with coverage allele-fraction above 0.05. We looked for regions where (i) SNP density within a window of 90–120 bp was high, (ii) the combination of SNPs was variable across the 75 potato lines sequenced, and (iii) this region was flanked by conserved sequence across the 75 potato lines. Those conserved sequences were targeted for primer design. We extracted the sequence of the targeted region with samtools faidx and used blastn (Altschul et al. 1990) to check for sequence similarity with off-target regions (only single-copy regions were retained). Primer 3 plus (Untergasser et al. 2012) was used for primer design with the following settings: product size 165-180 nt, primer size 15-(opt.25)-35 nt, Primer Tm 60-(opt.62)-65 C, 40-(opt.50)-65% GC, and the coordinates of the “Pair OK Region List” (start and stop of the flanking conserved sequences). Once a primer pair was successfully designed, we targeted a new region 1 Mb ± 0.1 Mb downstream of the previous target. We designed 10 additional primer pairs, using less stringent criteria, to target some disease-resistance markers routinely tested in the breeding program by kompetitive allele-specific PCR (KASP) (He et al. 2014) or retrieved from the literature (Yuan et al. 2020). A summary of the pipeline employed for primer design is illustrated in Figure 1.

Table 1. Euchromatic regions targeted by PotatoMASH. Number of core loci targeted for primer design.

Chr/arm	Start	End	Length (Mb)	Core Loci	Diagnostic Loci	Total Loci
chr1/1	1	6,236,423	6.2	6		
chr1/2	58,566,960	88,663,952	30.1	31		
chr2	18,620,376	48,614,681	30.0	31	2	
chr3/1	1	5,853,851	5.9	6		
chr3/2	37,557,548	62290286	24.7	25		
chr4/1	1	10,893,487	10.9	11	2	
chr4/2	50,527,797	72,208,621	21.7	24		
chr5/1	1	10,773,566	10.8	12	1	
chr5/2	42,795,302	52,070,158	9.3	11	1	
chr6/1	1	6,372,027	6.4	7	1	
chr6/2	37,792,178	59,532,096	21.7	22		
chr7/1	1	7,298,544	7.3	9		
chr7/2	36,698,521	56,760,843	20.1	21		
chr8/1	1	6,899,227	6.9	7		
chr8/2	35,611,618	56,938,457	21.3	23		
chr9/1	1	9,549,714	9.5	10		
chr9/2	44,754,712	61,540,751	16.8	18		
chr10/1	1	5,591,854	5.6	6		
chr10/2	47,231,005	59,756,223	12.5	14		
chr11/1	1	10,117,653	10.1	11	2	
chr11/2	35,737,669	45,475,667	9.7	11		
chr12/1	1	9,273,808	9.3	11		
chr12/2	50,482,591	61,165,649	10.7	12	1	
Total			318 Mb	**339**	10	347

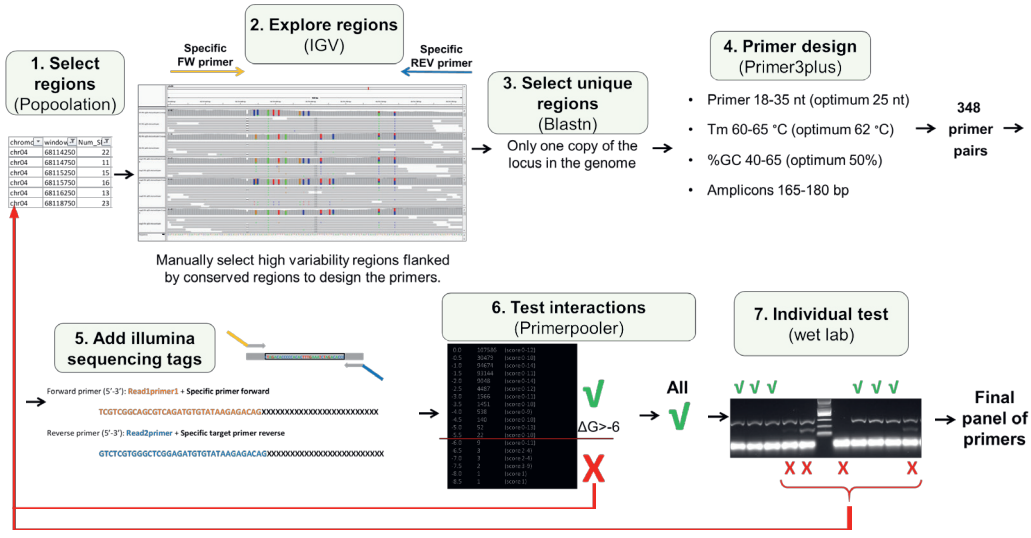


Figure 1. Pipeline for PotatoMASH primers design.

The method to construct multiplex amplicon libraries for Illumina sequencing was based on the GT-Seq method (Campbell et al. 2015). This method consists of an initial multiplex PCR with tailed specific primers. The resulting products include the selected regions to be sequenced flanked by the Illumina sequencing primer tags R1 and R2 (Figure 2). This product is then used as a template for a second PCR in which the Illumina sequencing adapters P5 and P7 are incorporated, a unique 6nt i7 barcode to identify the plate a sample originates from, and a unique 6nt i5 barcode to identify the sample within that plate. In order to achieve this, once all primers were designed, we added the tag for R1 Illumina sequencing primer at 50 extreme of each forward primer (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGFW primer) and the tag sequence for R2 Illumina primer at 50 end of each reverse primer (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-REV primer). Thus, the primers for the first PCR ranged from 51-69 nt in length.

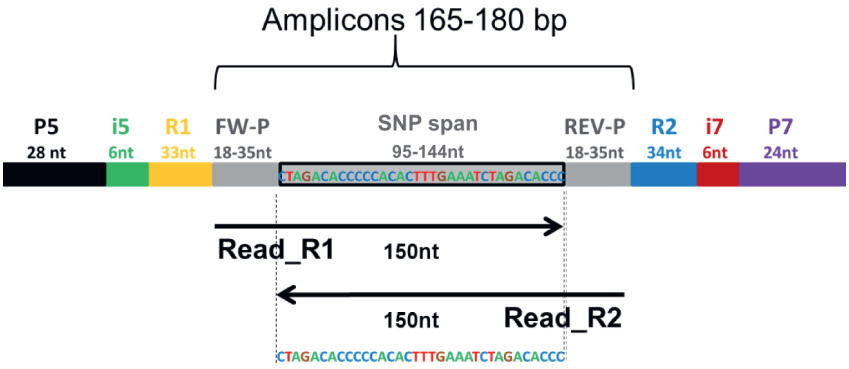


Figure 2. PotatoMASH library structure in relation to paired-end 150nt Illumina reads Read_R1 and Read_R2.

To avoid primer interactions and putative secondary products, all designed primers were tested in silico using primerpooler (Brown et al. 2017) with the following settings: T = 57 ◦C (the lowest annealing temperature that will be used), Magnesium (divalent cations) = 3 mM, ΔG threshold = −6 , and maximum amplicon length 2000 bp. Primer interactions with a ΔG lower than the threshold were replaced, and final primer sets were ordered from Integrated DNA technologies (IDT, Iowa, USA) at 750 µM each. Forward primers for the second PCR were composed of the Illumina adapter tag, a unique 6nt barcode (i5, up to 96 Truseq/NEB 6 base index), and the first 14 bases of R1 tag (AATGATACGGCGA CCACCGAGATCTACAC-i5-TCGTCGGCAGCGTC). They were ordered in a 96-well plate format at a concentration of 100 µM. We also ordered eight i7 reverse primers in tubes at 100 µM for the second tailed PCR. They were composed of the Illumina adapter P7, a unique 6nt barcode (i7, up to 8 Agilent SureSelectXT Custom kit index), and the first 15 bases of R2 tag (CAAGCAGAAGACGGCATACGAGAT—reverse complementary sequence of i7 index GTCTCGTGGGCTCGG). The number of possible pairwise combinations for this set of i5 and i7 barcodes is 768 samples but a higher number of samples can be processed with

additional i7 barcodes. These barcoding primers (i5-primers and i7-primers) were diluted individually to a working solution of 10 µM.

Each primer pair was tested individually with 40 ng of potato DNA, 25 nM each primer, 3 mM MgCl2, 40 µM each dNTP, high-fidelity Q5 polymerase (NEB M0491L) at 0.02 U/µL and Q5 enhancer (see below for PCR conditions). The second PCR was performed with the same mixture without Q5 enhancer but one i5-primer and one i7-primer at 1 µM each. The PCR products were visualized on 1.2% agarose gels. The expected size of the PCR products ranged between 297 and 312 bp. Any primer pair with low efficiency or producing secondary products (around 14 % of primer pairs) were replaced with alternative primers targeting the same or nearby region. The final selected primer pairs were pooled together by combining 2.5 µL of each of the 694 primers and diluted by adding 11.104 mL of ddH2O to a working concentration of 125 nM for each primer (250 nM/primer pair). A complete list of primer sequences used in PotatoMASH is included as Supplementary Materials (File S1).

Genotyping Panel

For this work, we used DNA from a collection of 705 potato lines referred to as the FRY population previously used in genetic analysis for tuber quality traits (Byrne et al. 2020). We also extracted DNA from 60 additional potato lines selected from the sixth year of the Teagasc/IPM breeding programme (TPBP_2020_Y6) using a GenElute™ Plant Genomic DNA Miniprep Kit (Sigma, G2N10, MA, USA). DNA was quantified using a Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen, P7589, MA, USA) and normalized to a concentration of 20 ng/µL. The 765 lines are referred to as the Extended FRY population.

PotatoMASH Library Construction

Libraries were constructed using a two-step PCR strategy (Figure 3). The cocktail for amplification of target loci (PCR1) included: 0.1 µL ddH2O, 1.4 µL pooled primer mix 125 nM each primer (final concentration is 25 nM each primer, 50 nM per primer pair), 3.5 µL of Qiagen Plus multiplex master mix (QPMMM, Qiagen, 206152, Hilden, Germany), and 2 µL of template DNA (40 ng).

PCR was carried out in a gradient thermocycler in 96-well PCR plates with the following conditions for PCR1: 95 ◦C 15 min; 8 cycles × (95 ◦C 30 s, 0.2 ◦C/s ramp down to 57 ◦C annealing 30 s, 72 ◦Cmin); 16 cycles × (95 ◦C 30 s, 65 ◦C 30 s, 72 ◦C 30 s); 10 ◦C hold. Following PCR1, the amplified samples were diluted 15-fold by adding 100 µL of ddH2O and mixed by pipetting up and down.

PCR2 adds indices that effectively identify each sample by well and by plate. A mix for each plate was made with 1 µL of 10 µM plate-specific i7-primer and 5 µL of QPMMM, and 6 µL of this PCR2 cocktail was added to each well. Next, 1 µL of 10 µM well-specific i5-primers and 3 µL of the diluted PCR1 product were added to the appropriate wells. PCR was conducted with the following conditions: 95 ◦C 15 min; 10 cycles × (98 ◦C 10 s, 65 ◦C 30 s, 72 ◦C 30 s); 72 ◦C 5 min; 10 ◦C hold.

Following PCR2, each plate of the libraries was normalized using the SequalPrep™ Normalization Plate Kit, 96-well (Applied Biosystems, A1051001, Waltham, MA, USA). This kit provides amplicon purification and normalization of PCR product concentration via a limited binding capacity of the solid-phase coating the walls of the plate wells. Following normalization, 15 µL of each sample per 96-well plate was pooled into one tube for a total of 8 tubes. A concentration-purification step was then performed on each of the tubes by mixing 7.5 mL of binding buffer (PB buffer, Qiagen, 19066) and using QIAquick PCR Purification Kit (Qiagen, 28104), following the manufacturer instructions. The product was eluted in 40 µL of elution buffer.

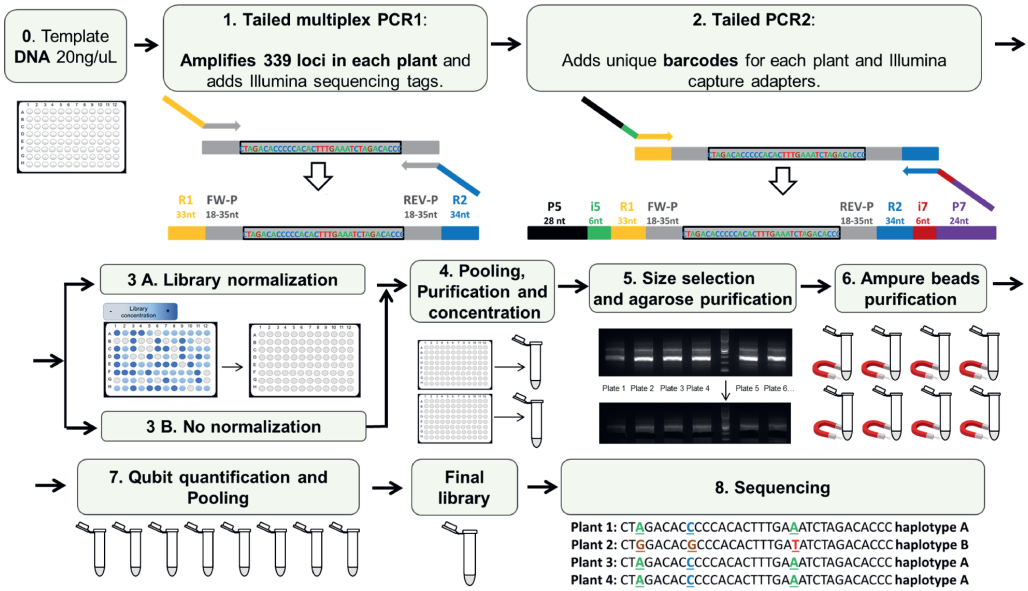


Figure 3. PotatoMASH library construction.

In contrast with the original GT-Seq protocol (Campbell et al. 2015), we normalized template DNA concentration at the outset, and we therefore tested the possibility of removing the library normalization step by sequencing the same sample set with and without library normalization (Figure 3-step 3B). In that case, we used the original number of cycles established by GT-Seq authors for PCR1: 95 ◦C 15 min; 5 cycles × (95 ◦C 30s, 0.2 ◦C/s ramp down to 57 ◦C annealing 30 s, 72 ◦C min); 10 cycles × (95 ◦C 30 s, 65 ◦C 30 s, 72 ◦C 30 s); 10 ◦C hold. We pooled 5 µL from each well after PCR2, took half volume of the pooled sublibrary for each plate (240 µL), used 1.2 mL of PB buffer for the concentration-purification step, and eluted it in 40 uL of elution buffer.

For all products after library normalization (Figure 3-step 3A) or without library normalization (Figure 3-step 3B), each 40 µL aliquot was run on 1.2% agarose gel. Gel slices containing the product around 300 bp were purified by using Wizard SV Gel and PCR Clean-Up System (Promega, A9281, Madison, WI, USA), and the product was eluted in 50 µL of elution buffer.

The last purification step was then performed on each of the aliquots by adding and mixing 25 µL (0.5× volume) of AMPure XP magnetic beads (Beckman Coulter, A63881, Brea, CA, USA). Each tube was then placed on a magnetic rack. The supernatant was transferred to a fresh tube and mixed with 75 µL (0.6× volume) of magnetic beads and placed in the magnetic rack. The supernatant was discarded, and the immobilized beads were washed with 180 µL of 85% ethanol. Purified libraries were then eluted with 30 µL of nuclease-free ddH2O and transferred to fresh 1.5 mL tubes.

Following purification, each of the 8 plate libraries were quantified using a Qubit™ dsDNA BR Assay Kit (Invitrogen, Q32853, Boston, MA, USA), and equal molecular amounts were pooled to create the final library for sequencing. The final library containing 765 individuals was sequenced mixed with 50% PhiX on one lane of Illumina HiseqX instrument by Novogene (Cambridge, UK) Company Limited to obtain paired-end 2 × 150 nt reads. Fastq files are available in the BioProject database under BioProject ID PRJNA858449. More detailed information about how to perform PotatoMASH can be found at https://doi.org/10.17504 /protocols.io.e6nvw53zdvmk/v1 (accessed on 1 September 2022).

Multiallelic Haplotype Analysis

Fastq files were de-multiplexed, barcodes removed, and read pairs merged using FLASH (Magoč and Salzberg 2011) (min overlap -m 50; Max overlap -M 150). We filtered the merged reads using the fastx toolkit (Gordon 2010) (minimum base quality (-q30) in 90% of bases (-p90)). Merged and filtered reads were then mapped to the *S. tuberosum* genome v4.04 (Hardigan et al. 2016) with BWA-MEM (Li 2013). Variant calling was performed with bcftools (Li 2011), using the classic (biallelic) model: bcftools mpileup -Ou -I –max-depth 8000 -min-MQ 30 -a DP,AD -f potato_dm_v404_all_pm_un.fasta -b bam.list | bcftools call -cv -Ob -f GQ -o PotatoMASH.bcf. Then we filtered high quality SNPs with vcftools: vcftools –bcf PotatoMASH.bcf –out PotatoMASH –min-alleles 2 –max-alleles 2 –recode –recode-INFO-all –minQ 30 –minDP 6 –maf 0.05 –max-maf 0.95 –remove-filtered-all –maxmissing 0.5. For the Normalized library (Figure 3-step 3A), out of 5348 SNPs identified during SNP calling, 2236 were filtered based on minimum coverage 6 and min mapping quality 30. For the non-normalized library (Figure 3-step 3B), we filtered 2279 SNPs out of 5104 sites. SMAP (Schaumont et al. 2022) *haplotype-sites* v4.1.1 was run with the following parameters -read_type merged -partial exclude –no_indels –discrete_calls dosage –frequency_interval_bounds 12.5 12.5 37.5 37.5 62.5 62.5 87.5 87.5 –dosage_filter 4 –min_read_count 20 –min_haplotype_frequency 5 –min_distinct_haplotypes 0. SMAP *haplotype-sites* requires the loci coordinates that can be calculated by mapping the primers (Supplementary Materials File S1) to the genome. We included the bed file for potato genome DM_v4.04, which was used in this work, and other bed files for different versions of SMAP and potato genome DM_v6.1 (Supplementary Materials File S2). The output “haplotypes_discrete_calls_filtered” table (Supplementary Materials File S3) was used for downstream analysis.

On the other hand, the allele frequencies for the 10 different diagnostic SNPs were extracted from the original vcf file before filtering. A minimum of 20 reads was required. Dosage calls were calculated according to the % of reads representing the alternative allele:

<12.5% = “0”; ≥12.5–37.5% = “1”; ≥37.5–62.5% = “2”; ≥62.5–87.5% = “3”; >87.5% = “4”. In order to detect the haplotype containing the diagnostic SNP, which dosage calls should be concordant with the SNP dosage and also to detect putative linked haplotypes in the core loci, SMAP output was loaded in Microsoft Excel (Microsoft Corporation). The loci containing the position of each diagnostic SNP and the loci nearby (up to 4 Mb upstream and downstream) were arranged so the 765 potato lines were shown in rows and the short multiallelic haplotypes in columns to be sorted by the SNP dosage.

Haplotype-Based GWAS to Identify QTL Associated with Fry Colour

In order to compare the discriminatory power of SNPs versus multiallelic haplotypes, we performed a GWAS analysis on a subset of 279 lines of the Extended FRY population, for which QTLs for the trait fry colour had previously been detected using ~40 k GBS-derived SNP markers (Byrne et al. 2020). Analysis was performed with both the 2279 filtered SNP set obtained from non-normalized library (Figure 3, step 3B) and the 2012 multiallelic haplotypes detected by SMAP out of these 2279 SNPs.

The phenotypic data for fry colour ‘off-the-field’ (OTF) were generated in the Teagasc/IPM breeding program in 2017 (Byrne et al. 2020). GWAS was carried out with the R package GWASpoly (Rosyara et al. 2016). Haplotypes were treated as ‘Pseudo SNPs’ by effectively rating each individual haplotype as a biallelic presence absence marker, with presence indicated by 1, 2, 3, or 4 depending on dosage and absence coded as 0. Each individual haplotype allele was assigned a different position within the locus region so that GWASpoly could handle the input file with allele dosage information (Supplementary Materials File S4).

Population structure was controlled using the K model, where the covariance matrix was calculated using all SNPs, and QQ plots were used to assess if there was sufficient control of population structure (QQ-Plots in Supplementary Materials Figure S5a). The function GWASpoly with an additive model was used to test for association at each marker. Instead of filtering markers based on minor allele frequency, the maximum genotype frequency option was used (geno.freq = 1–10/279), so haplotypes present in fewer than 10 individuals were removed. The genome-wide false discovery rate was controlled using Bonferroni correction (at a significance level of 0.05).

Mapping Population Genotyping and Linkage Map Construction

The diploid potato population FRW19-112 was developed from a cross between the *S. tuberosum* clone RH89-039-16 (Zhou et al. 2020) and the breeding clone bearing a *S. microdontum* and *S. tuberosum* ancestry IVP10-281-1 (Meade et al. 2020c). A population of 92 FRW19-112 plants was grown from true seeds in five-litre pots in an open ground greenhouse compartment with drip irrigation and a wet pad-and-fan evaporation cooling

system. Young leaf material from the parental clones and the population was collected in 96 deep-well plates and freeze-dried for 48 hours prior to genomic DNA extraction.

DNAs were extracted with Mag-Bind® Plant DNA DS 96 Kit (Omega-VWR M1130-00, Philadelphia, USA), quantified, and normalized to 20 ng/µL as described in Section 2.2. Libraries were constructed as described in Section 2.3 with library normalization (Figure 3, step 3A). The final library containing 94 individuals was sequenced on an Illumina Novaseq 6000 instrument by Novogene (UK) to obtain paired-end 2 × 150 nt reads.

Fastq files are available in the BioProject database under BioProject ID PRJNA858449.

The pipeline to obtain the short multiallelic haplotypes was the same as described in Section 2.4. Out of 1805 SNPs identified during SNP calling, 1289 were filtered based on minimum coverage 6 and min mapping quality 30. SMAP *haplotype-sites* v4.1.1 was run with parameters -read_type merged -partial exclude –discrete_calls dosage –frequency_interval_bounds 10 10 90 90 –dosage_filter 2 –min_read_count 10 –min_haplotype_frequency 20 –locus_correctness 90. The output "haplotypes_discrete_calls_filtered" table containing 844 haplotypes was used for downstream analysis.

Prior to map construction, three F1 clones with more than 10% missing haplotypes were removed, and ten haplotypes, for which one of the parental dosages was missing, were imputed based on the observed offspring segregation. The best-fitting segregation model of each short multiallelic haplotype was identified using the function CheckF1 of polymapR (Bourke et al. 2018). SMAP haplotypes were further filtered with the removal of 26 haplotypes with missing and unimputable parental dosages of 59 strongly distorted haplotypes and of 54 non-segregating haplotypes, resulting in 705 retained haplotypes (Supplementary Materials File S6). Subsequently, 154 haplotypes showing identical segregation patterns with at least one other haplotype were binned, yielding 551 uniquely segregating haplotypes and haplotype bins for the linkage map construction.

Chromosomal linkage maps were constructed using polymapR version 1.1.2 following the package vignette with minor modifications to fit our diploid data. Pairwise estimators for recombination frequency and their associated LOD scores were determined for all multiallelic haplotypes and clustered based on their LOD scores. Twelve chromosomal clusters were identified at a LOD score threshold of 4.5. Cluster numbers were replaced with DMv4.04 chromosome numbering for consistency with the physical map used during read alignment. Next, haplotypes were ordered, and an integrated linkage map was created using MDSmap_from_list, a wrapper function around the estimate.map function from MDSMap (Preedy and Hackett 2016). During the mapping process, two non-clustering and seven outlying haplotype bins with a high nearest-neighbour fit score or an abnormal position in the principal curve analysis were removed. The haplotypes that were binned because of their identical segregation patterns were then added back to the map, resulting in 690 mapped haplotypes. PolyoriginR version 0.03 (Zheng et al. 2021) was then used to phase the haplotypes into parental homologs with a recombination rate per chromosome set at

1.25. The output was converted back into polymapR format to be visualized with the function plot_phased_maplist.

Results

Potato Multi-Allele Scanning Haplotags (PotatoMASH) as a Genotyping System

We multiplexed, in a single PCR reaction, 339 loci placed at equal spacing throughout the gene-rich portion of the 12 chromosomes of potato (Figure 4). Figure 4 represents the positions of the PotatoMASH core loci covering the euchromatic portion of the genome flanking the centromeric heterochromatin, except chromosome 2. Chr 2 is acrocentric, and the short arm is composed of the nucleolar organizing regions within the heterochromatin. Therefore, we only selected regions in the long arm.

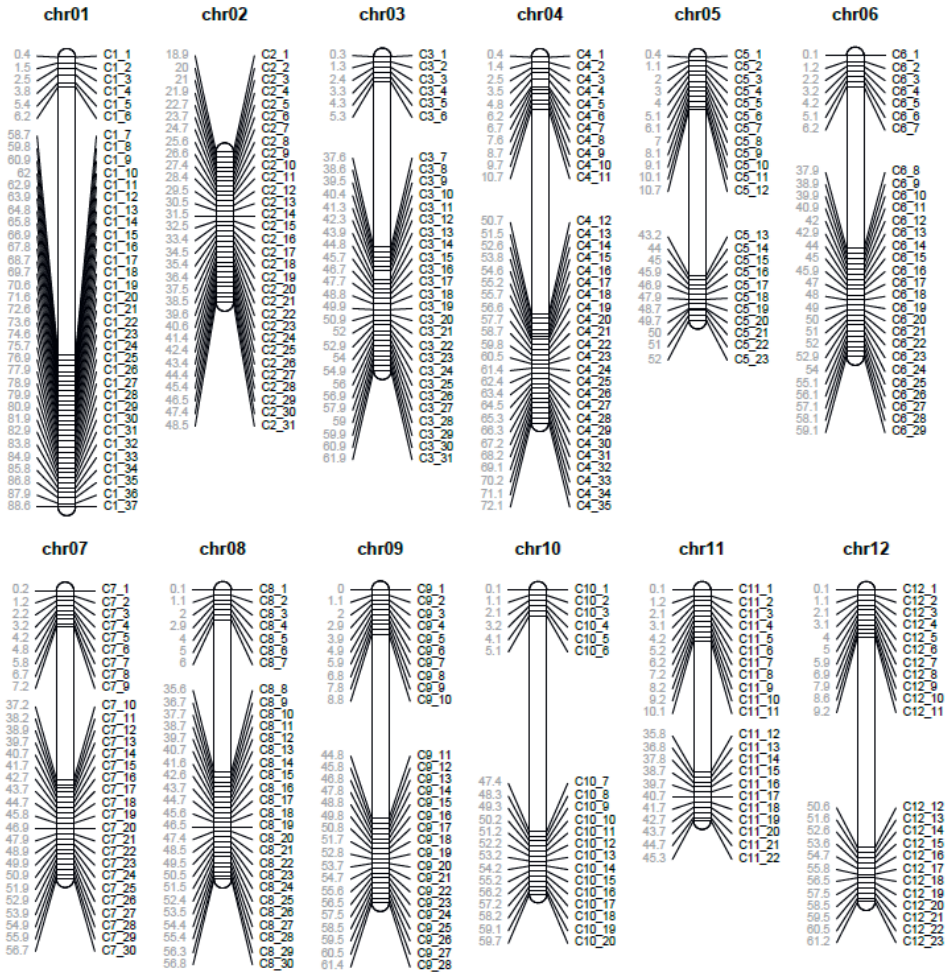


Figure 4. The physical map of potato based on pseudochromosome molecule assembly of the DM_v4.04 reference sequence. The positions of the 339 PotatoMASH core loci are represented in Mb intervals.

To test the ability of the primer set to reveal allelic diversity in tetraploid potato breeding germplasm, we tested PotatoMASH initially in the Extended FRY population. Normalization of samples prior to sequencing is a major cost component of the GT-Seq process as originally described (~20% of the per-assay cost), so we performed the experiment twice, sequencing both normalized and non-normalized libraries in order to test whether the normalization step could be left out, considerably cheapening the assay.

We obtained 56.6 Gb of sequencing data distributed across all 765 potato samples for the normalized library (theoretically ~0.5 M raw reads or 0.25 M paired-end reads/sample, 700 paired-end reads/locus) and 56.1 Gb for the non-normalized library. We detected some low-output samples (less than 36,000 raw reads/sample, less than 50 paired-end reads/locus) corresponding to two batches of samples distributed between plates 6, 7, and 8 (Figure 5). As expected, the number of low-output samples was lower in the normalized library (32 samples) than in the non-normalized library (81 samples). On the other hand, library normalization led to an enrichment of amplicons of the most efficient primer pairs leading to lower read depth at other loci (Figure 5a). We did not observe the same problem with the non-normalized library (Figure 5b). Therefore, in this study, normalization introduced more variability in the coverage per locus (Figure 5). After merging and filtering reads, we retained 228,420 reads/sample on average. The efficiency of the primer pairs (either low or high) was consistent across all samples (Figure 5), which indicates that the amplification efficiency of the primers is not genotype-dependent.

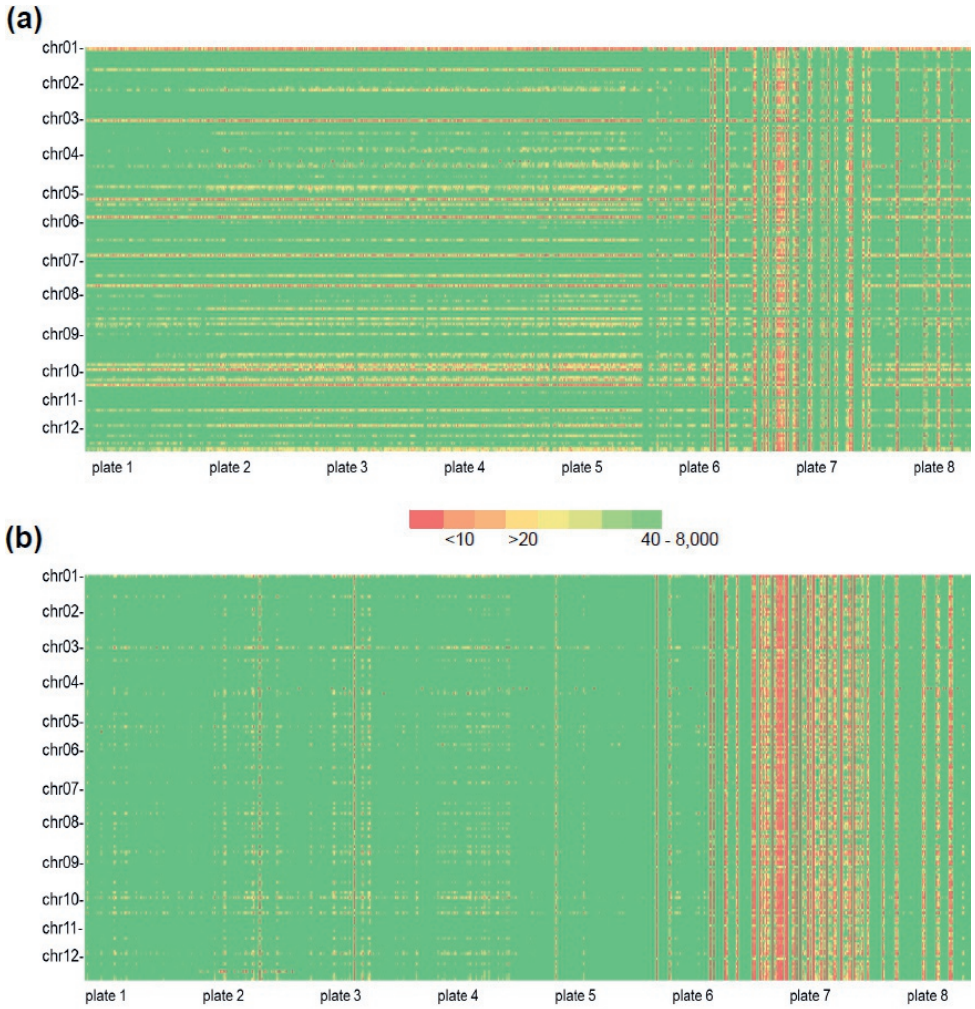
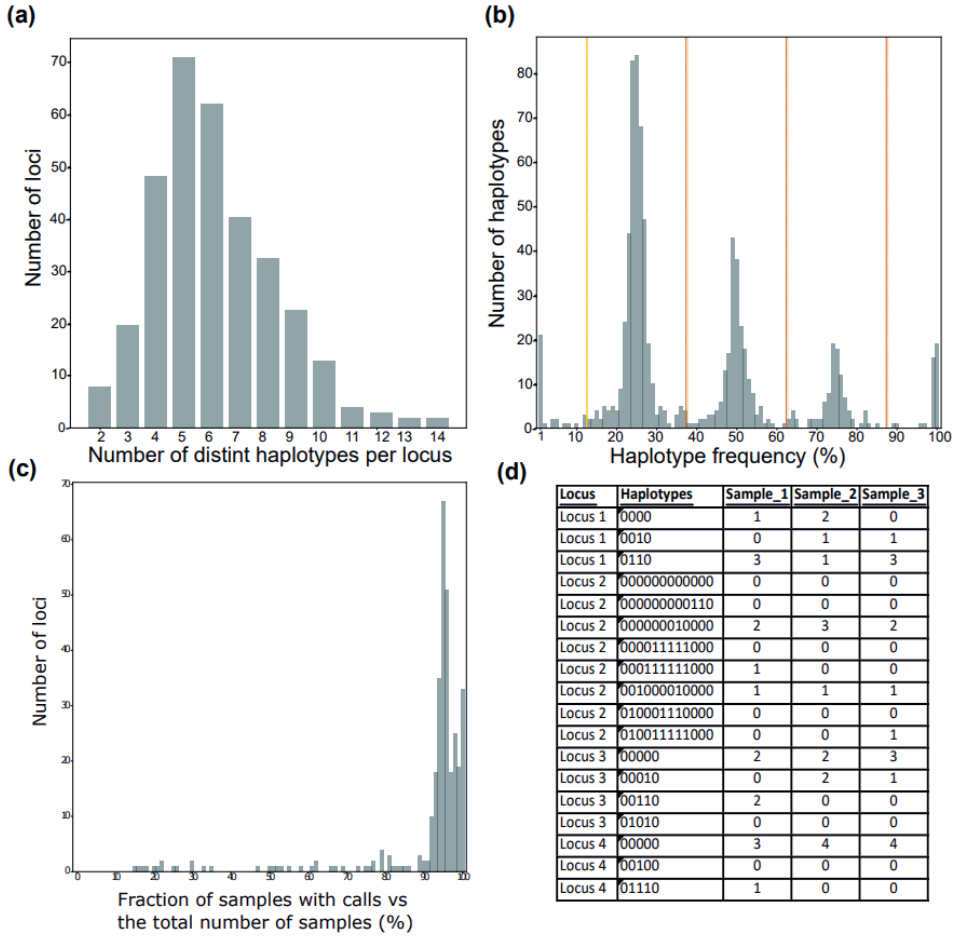


*Figure 5. Coverage of the 339 PotatoMASH core loci. Heat map of the number of merged and filtered reads of 765 samples (in columns) that mapped to each locus (in rows). (**a**) Normalized library (**b**) Non-normalized library.*

After filtering, the normalized and non-normalized libraries revealed 2236 and 2279 SNPs loci, respectively. For the normalized library, of the 339 loci, 333 yielded haplotype data when the SNP dataset was processed with SMAP *haplotype-sites*. Six loci produced reads, but for various reasons, they failed to generate haplotype information. We obtained a total of 2032 short multiallelic haplotypes across the population in the remaining 333 core loci, ranging from 2–14 haplotypes per loci, whilst most loci showed 5–6 haplotypes. The four alleles for each locus/sample were successfully detected in 84% of sites (locus/sample), and the rest are reported as NA. The majority of loci obtained calls for more than 90% of samples.

In the non-normalized library, five of the six failed loci observed in the normalized library were considered non-polymorphic, and we obtained a total of 2012 multiallelic haplotypes across the population in the other 334 core loci, ranging from 2–14 haplotypes per loci, whilst most loci showed 5–6 haplotypes (Figure 6a). The four alleles for each locus/sample were successfully detected in 84% of sites. The haplotype call frequency for each individual showed a tetraploid haplotype frequency distribution profile (Figure 6b).

The majority of loci got calls for more than 80% of samples (Figure 6c). As final output, we obtained a table with discrete dosage calls for each haplotype in each sample (Figure 6d), which was used for downstream analysis.



Locus	Haplotypes	Sample_1	Sample_2	Sample_3
Locus 1	0000	1	2	0
Locus 1	0010	0	1	1
Locus 1	0110	3	1	3
Locus 2	000000000000	0	0	0
Locus 2	000000000110	0	0	0
Locus 2	000000010000	2	3	2
Locus 2	000011111000	0	0	0
Locus 2	000111111000	1	0	0
Locus 2	001000010000	1	1	1
Locus 2	010001110000	0	0	0
Locus 2	010011111000	0	0	1
Locus 3	00000	2	2	3
Locus 3	00010	0	2	1
Locus 3	00110	2	0	0
Locus 3	01010	0	0	0
Locus 4	00000	3	4	4
Locus 4	00100	0	0	0
Locus 4	01110	1	0	0

*Figure 6. (a) Haplotype diversity distribution of 333 loci across the 765 individuals in the dataset generated by the non-normalized library. (b) Haplotype frequency spectrum of one individual cv. Gravity. (**c**) Locus call*

Demonstrating the Expandability of the PotatoMASH Platform Using Targeted R-Locus Markers

In addition to the 339 core loci, to demonstrate the ability to add markers with specific targets to PotatoMASH, we also designed primers to capture SNPs linked to disease and pest resistance loci of interest to the Teagasc/IPM Potato Group breeding programme. Data for 10 loci involved in resistance to common scab, the potato cyst nematodes *Globodera pallida* and *G. rostochiensis*, late blight, potato virus Y, and potato wart disease are shown in (Table 2). All of the target loci were successfully amplified, and the target SNPs were detected in the original vcf file before filtering. However, three loci did not generate a haplotype associated to the target SNP subsequent to data processing by SMAP *haplotype-sites*. In the other cases, the concordance between the dosage diagnostic SNP and the haplotype was close to 100%.

Table 2. Multiallelic haplotypes linked to disease resistance markers: Ref. = Literature reference/Developed by Teagasc Potato Breeding Program (TPBP); Locus = Name of the locus containing the marker in PotatoMASH; Haplotype = haplotype containing target SNP; Concord. = % concordance between marker dosage and haplotype dosage.

Ref.	Resistance to:	Name	SNP Position	Locus	Haplotype	Concord.
(Yuan et al. 2020)	*S. scabies*	c2_17867	chr02:36548178 [T/C]	C2_B2	011110	99.9%
(Yuan et al. 2020)	*S. scabies*	c2_17864	chr02:36550070 [T/C]	C2_B3	0011110	99.9%
TPBP	*G. pallida (Pa2/3)*	Gpa4	chr04:4782401 [A/G]	C4_5	001110	99.9%
(Meade et al. 2020a)	*P. infestans*	R2	chr04:6191864 [A/T] chr04:6191873,76,77 [TGATT/CGAAA]	C4_6	Not detected	NA
(Rouppe van der Voort et al. 2000)	*G. pallida*	Gpa5	chr05:5485534 [T/A]	C5_B9	01101010101111 00101	96.5%
(Meade et al. 2020a)	*G.rostochiensis (P1/4)*	H1	chr05:49238169 [T/A]	C5_B10	000000110	100%
TPBP	*P. infestans*	Rpi-blb2	chr06:775752 [G/A]	C6_B1	Not detected	NA
(van Eck et al. 2017)	PVY	Ny(o,n)sto	chr11:284162 [T/C] 68 [T/C]	C11_B1	000101010010	98.95%
(Prodhomme et al. 2020)	*S. endobioticum*	Sen1	chr11:3928601 [A/G]	C11_B3	001100	100%
(Grech-Baran et al. 2020)	PVY	Ry-sfto	chr12:59957417 [G/A]	C12_B6	Not detected	NA

Haplotype-Based GWAS to Identify QTL Associated with Fry Colour

The design of PotatoMASH combines even marker spacing across the euchromatic portion of the genome and the ability to reveal multiple haplotypes at each locus to efficiently scan genome-wide variation. One of the main applications for this is in genetic marker discovery. We tested the ability of the haplotypes and the SNP set from which they were derived (in the non-normalized dataset) to discover QTL on chromosomes 10 and 2 for fry colour that

had previously been detected in a portion of the FRY population using >40 k GBS-derived SNP markers. We did not identify any significant QTL using the 2279 biallelic SNPs underlying the haplotypes (Figure 7a).

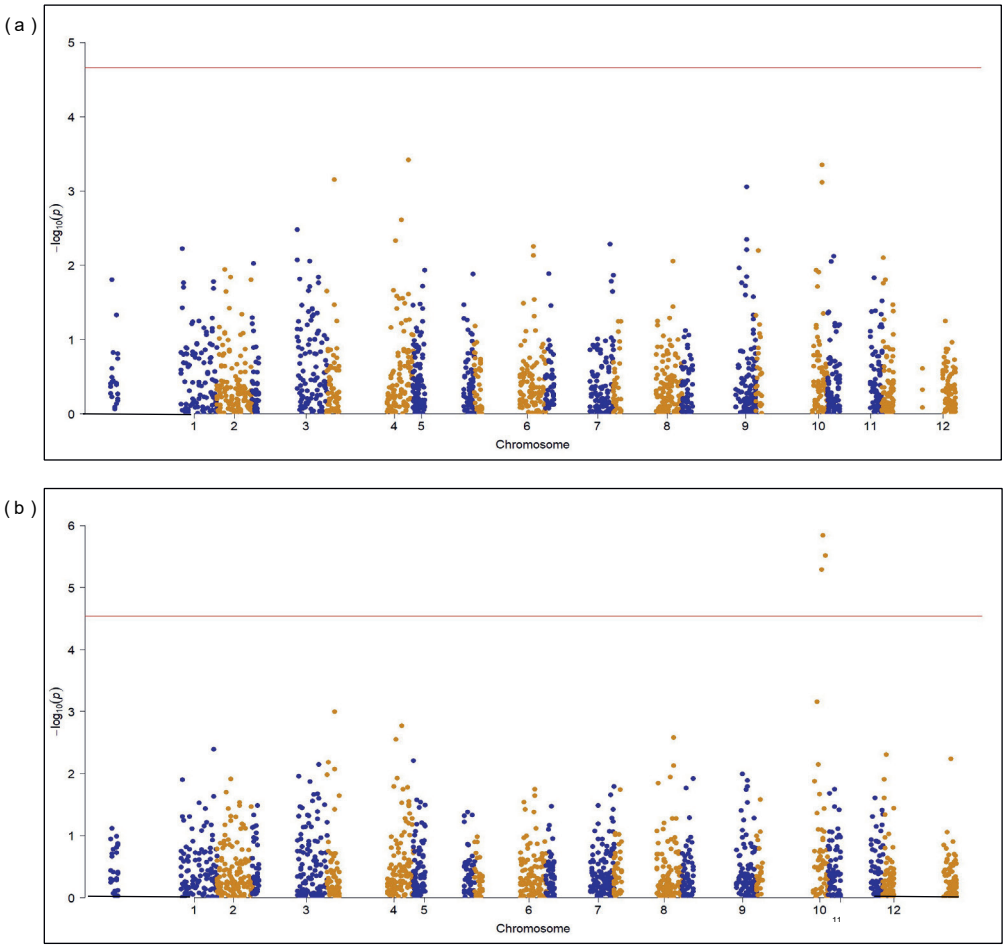


Figure 7. Manhattan plot of GWAS results, additive model, for 'off-the-field' fry colour (OTF) in 2017 population with the SNPs (a) and the multiallelic haplotypes (b) Horizontal line shows the QTL significance threshold at 4.54 (Bonferroni correction, level = 0.05).

However, amongst the 2012 haplotypes, which are 2012 different combinations of 2279 biallelic SNPs, three haplotypes were significantly associated with fry colour on chromosome 10 (Figure 7b). The three associated haplotypes underlying the QTL, namely, 10_14_00000100100 (Chr10:54209422-54209550), C10_15_01101010 (Chr10:55208521-55208635), and C10_17_000110 (Chr10:57186478-57186604), were carrying two, four, and two SNPs in loci C10_14, C10_15, and C10_17 respectively. These haplotypes showed the same segregation pattern and the presence of the haplotypes had a negative impact on fry colour (Boxplots in Supplementary Materials Figure S5b).

These results agree with those obtained by Byrne et al. (2020) that analysed the same population using a marker set consisting of 46,406 SNPs generated using GBS. In that study, a QTL on chr10 was identified with a large cluster of associated SNPs between 49 and 59 Mb, peaking at 55.28 Mb. Our results are consistent with those findings, albeit using a much lower-density marker set. We did not detect the QTL on chromosome 2, which was marginal in the original study. We also tracked the parental origin of the haplotype with the largest effect on fry colour (C10_14_00000100100), and for 27 out of 28 lines carrying this haplotype, the variety “Valor” was used as a parent or grandparent. Additional information about the SNPs composing the haplotypes identified in the Extended FRY population can be found in Supplementary Materials Table S8.

Linkage Map Construction Using PotatoMASH Haplotypes

We also applied PotatoMASH to a bi-parental diploid mapping population (FRW19112), both to test its performance in genetic mapping and to validate certain features of the assay. We obtained 20 Gb of sequencing data for the 94 individuals of the population (0.7 M paired-end reads/sample, 2000 PE reads/locus). After merging and filtering reads, we retained 484,134 reads/sample on average (1428 reads per sample/locus). After multiallelic haplotype analysis, we obtained a total of 844 haplotypes across the population in 309 core loci (2.7 haplotypes/locus), ranging from 1–4 haplotypes per locus, whilst most loci showed three haplotypes. With the exception of two triploid clones, which were identified because of their haplotype frequency distribution profile, the haplotype call frequency for each individual showed the expected diploid profile (Supplementary Materials File S7).

The rate of missing data was extremely low; the two alleles for each locus/sample were successfully detected in 96% of sites, and the majority of loci obtained calls for more than 95% of samples. The SMAP *haplotype-sites* output table with dosage calls for each haplotype in each sample was further filtered as described in Section 2.6. A total of 690 short multiallelic haplotypes could be ordered and phased on 12 chromosomal linkage groups (Figure 8).

RH89-039-16 contributed with 274 female-specific haplotypes, while IVP10-281-10 contributed with 282 male-specific haplotypes. In addition, 134 haplotypes were segregating from both sides. The haplotypes were distributed relatively evenly across the 12 chromosomal linkage groups identified, with an average of 57.5 haplotypes per linkage group. However, more male than female haplotypes were discarded during the curation step because of the strong male-specific transmission ratio distortion on chromosome 1 and 12. This resulted in paternal linkage groups 1 and 12 composed of less haplotypes than their maternal counterpart. The total genetic map length was 880 cM, ranging from 53 cM for linkage group 6 to 95 cM for linkage group 1.

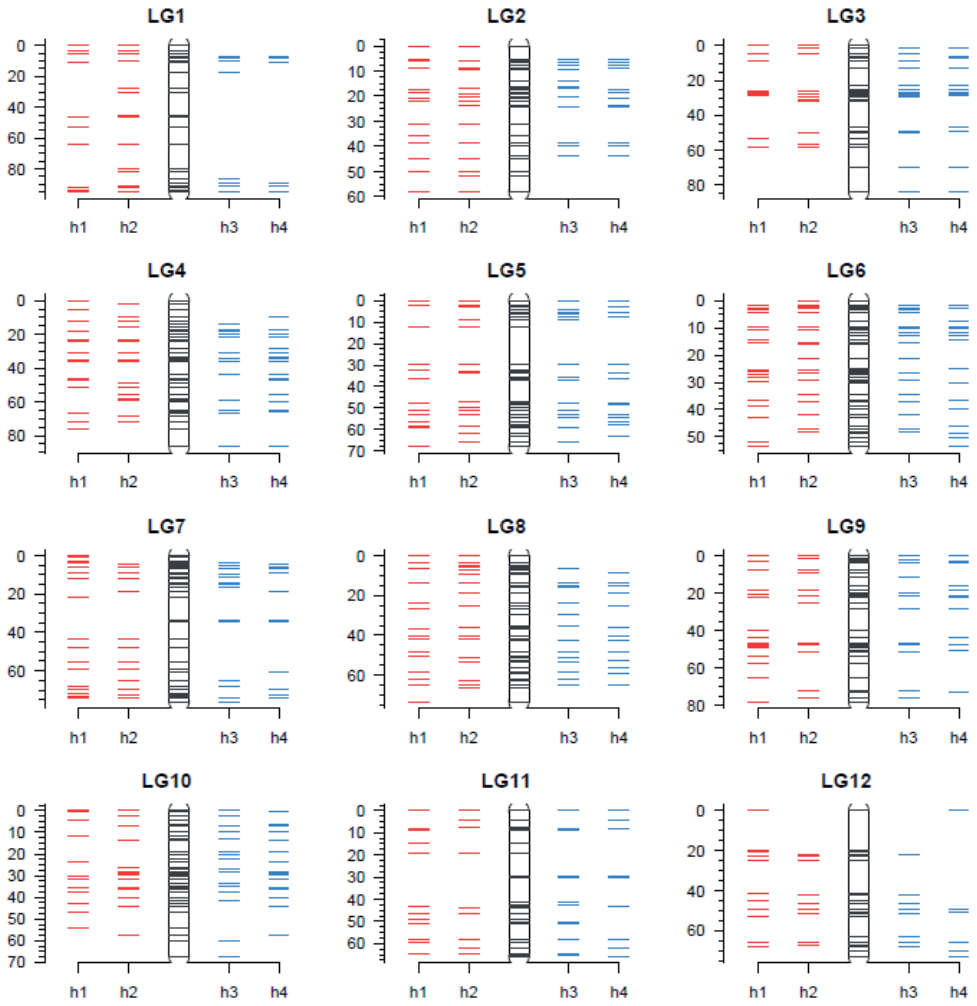


Figure 8. Phased homologue-specific map of population FRW19-112. Homologue maps from the diploid parent RH89-039-16 are shown in red (h1–h2), blue for IVP10-281-1 the diploid parent (h3–h4), and the integrated chromosomal map is shown in black.

The expected co-linearity between physical distance and genetic distance was observed for all linkage groups with the exception of few outliers, notably on linkage group 3 (Figure 9). Interestingly, this loss of co-linearity in linkage group 3 co-localizes with a 5.8 Mb paracentric inversion on the long arm of chromosome 3 recently identified among other clones in RH89-039-16, the female parent of our population (Tang et al. 2022). The lack of markers visualized as gaps in the long arm of chromosome 1 and on chromosome 12 coincide with the positions of markers discarded during the curation step due strong transmission ratio distortion. Surprisingly, gaps were also observed in regions not affected by this transmission ratio distortion, such as the short arm of chromosome 5 and the long arm of chromosome 7.

Such gaps could be due to the integration of female and male genetic maps with potentially different structures and recombination rates. On the other hand, we detected a linkage between the loci flanking the pericentromeric region, with an average genetic distance of 7.4 cM, despite an average per-chromosome physical distance of 46.7 Mb. This partially validates our initial premise that marker coverage in the pericentromeric heterochromatin was not required due to low levels of LD decay in this portion of the genome and indicates that our estimates of the heterochromatin-euchromatin border were sufficiently accurate for genome scanning purposes.

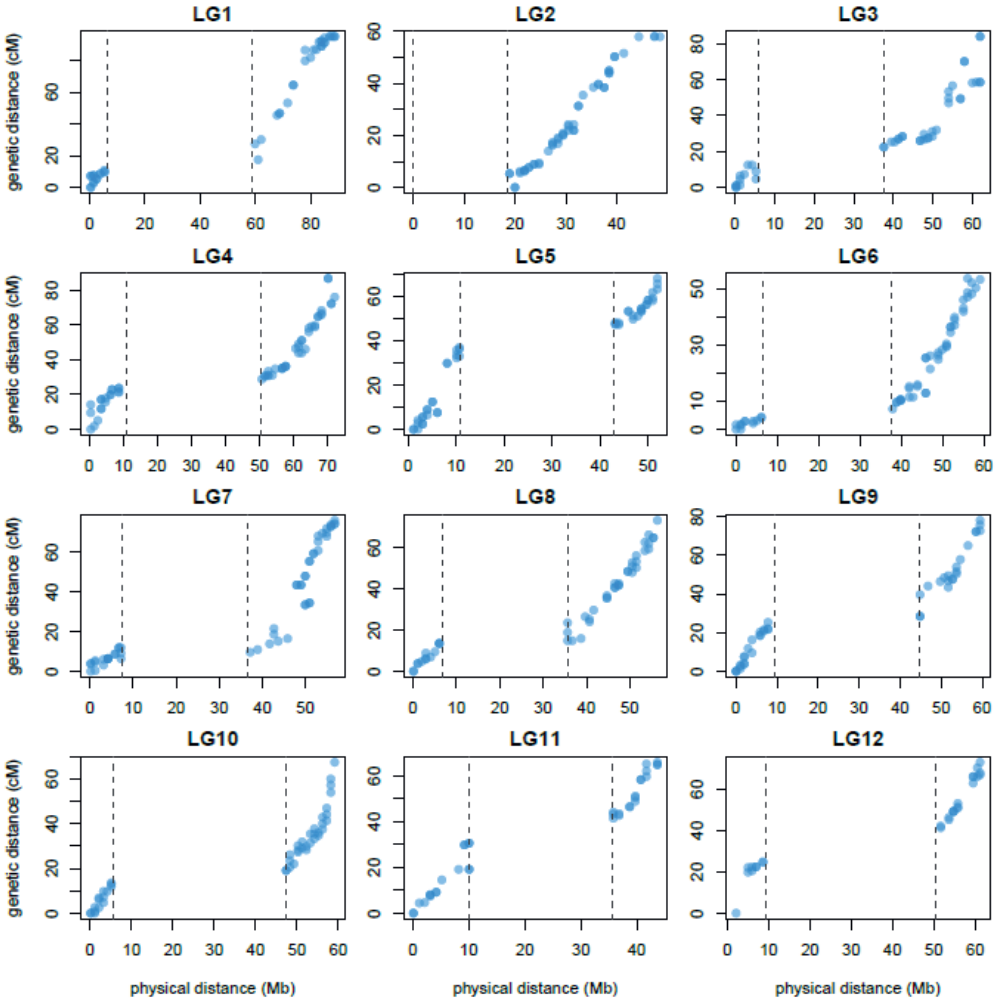


Figure 9. Plot of the genetic location (cM) vs. physical position (Mb) of PotatoMASH loci across the chromosomes. The vertical dotted lines represent the centromeric regions of the chromosomes.

Discussion

Coverage and Allelic Diversity

Marker coverage in terms of density and distribution across the genome is a major decision in the development of any genotyping platform. In their design of the SolSTW 20 k SNP array, Vos et al. (2017) posited that, based on adopting an LD decay threshold of r2 = 0.1, full SNP coverage of the haploid genome could be achieved by a minimum of 200 SNPs targeted at 2Mb intervals throughout the ~400 Mb euchromatic portion of the genome. However, this value would need to be upwardly adjusted to account for the number of haplotypes present; assuming 10 haplotypes per locus, this figure increases to 2000 SNPs. They also pointed out that an LD threshold of 0.1 is unlikely to have the ability to detect all QTLs. Based on a near total lack of LD decay between adjacent SNPs at 100 kb, a more comprehensive system would require the ability to survey 40,000 SNPs to ensure that at least one SNP was in LD with any other allele, including target QTL alleles. As our goal was to converge cost of application with a reasonable level of effectiveness, when designing PotatoMASH, we utilised the scenario with significant LD as a starting point. Depending on the threshold used to estimate LD decay, few studies give estimates of LD decay lower than ~0.5 Mb, so we adopted a 1 Mb spacing to ensure that no region was more than this distance away from a core locus. Using the euchromatin/heterochromatin boundaries described in the methods, this yielded 339 core loci, which were as close to this spacing as was achievable.

Whilst 1 Mb marker spacing allows physical coverage of the genome, it does not deal with high level of haplotypic diversity in potato, and we decided to adopt the “tag-level haplotype” concept of Tinker et al. (2016) to deal with this, using SMAP *haplotype-sites* (Schaumont et al. 2022) to provide a robust pipeline to identify short haplotypes. Within the Extended FRY population of 765 tetraploid individuals, we observed from 2–14 haplotypes per locus, with the majority of the loci exhibiting five or six haplotypes (Figure 6a), an average of 712 haplotypes per genotype and 2.4 (min 1.3–max 3) distinct haplotypes per genotype/locus. One relevant question is how close the PotatoMASH is to revealing the true full allelic or haplotypic diversity at these loci. As outlined earlier, allele copy number in European breeding germplasm, at least in genic regions, seems to span the range of 5–20 copies. Johan Willemsen (2018) estimated that in a panel of 83 tetraploid varieties, surveyed over 800 genic loci, an average of 25 haplotypes were present when 25 SNPs were aggregated over windows of ~500 nucleotides (Willemsen 2018). In reality, it is probable that PotatoMASH is underestimating the number of alleles per locus. Some of this is due to a combination of experimental and analytical pipeline features, in which features such as read depth/locus coverage and filtering, both within the SNP calling/filtering pipeline and SMAP *haplotype-sites*, result in some real SNPs not being used to discriminate between haplotypes. There is evidence in potato populations that many haplotypes are present at a low frequency due to their recent introduction. Potato is known to carry a high number of low-frequency SNP alleles (<1%), and this probably translates into a high level of low-

frequency haplotypes. Taking into account that we filtered the SNPs with minimum 5% of allele frequency prior to haplotype analysis, that we set up SMAP *haplotype-sites* to consider only haplotypes represented by at least 5% of the reads, and that SMAP *haplotype-sites* do not consider the positions of the indels, many real haplotypes could have been filtered out. In addition, as with all PCR-based genotyping approaches, there is also the possibility of null alleles arising from poorly or non-binding PCR primers due to sequence divergence in these regions (although we did take specific steps to mitigate this in the design phase). Larger structural variants (longer range presence–absence variation) could also cause null alleles. Finally, a haplotype is partially defined by the window of observation. In our case, the window of observation was 97–172 nucleotides, so it is conceivable that some haplotypes exhibiting identity in this window, especially when relatively few variants are observed, are flanked by polymorphic content that would split them into further haplotypic variants.

Performance of PotatoMASH across the Core Loci

In addition to cost (see below), another reason why we tried to minimise the number of core loci was to enable a high degree of manual intervention during the primer-pool design process, e.g., several-fold more regions than primers were individually manually inspected. An iterative process to attempt to minimise inter-primer-set interaction was adopted, and all candidate primers were individually tested in an attempt to maximise the per-locus success rate. Subsequently, 333 of the 339 core markers in the current PotatoMASH pool yielded SNP/haplotype information, a drop-out rate of only ~1.5%. In addition to the Extended FRY population, we also tested PotatoMASH on a diploid mapping population of 92 F1 individuals derived from two highly heterozygous parents. In contrast to the Extended FRY population, in the diploid mapping population, haplotypes were derived from loci, but 27 loci were non-polymorphic and 3 did not amplify (~10% drop-out rate). However, we have subsequently tested PotatoMASH across thousands of plants representing an even wider pool of material, and the core marker performance has remained stable (data not shown), with similar efficiency rates as in our tetraploid population (1.5% drop-out rate), demonstrating the utility in investing this time in the primer design phase. Because of the non-fixed nature of the primer pool, loci that consistently do not perform across experiments can be replaced with alternative primer sets in future applications of the PotatoMASH platform.

On the other hand, the diploid mapping population was partly an application oriented test—PotatoMASH could radically reduce the cost of diploid linkage mapping, and we wanted to test its effectiveness for this purpose. We were able to assemble a completely phased linkage map covering all four parental homologues across the 12 chromosomes of potato. As expected, low recombination rates in the pericentromeric heterochromatin meant that, despite the complete absence of markers in the area, markers flanking these regions exhibited linkage, with genetic distances ranging from 0.52 to 16.7 cM (average of 7.4 cM). Thus, our estimates of the heterochromatin/euchromatin boundaries, whilst somewhat arbitrary, seem reasonable. However, centromeric regions are not devoid of recombination,

and indeed, recombined centromeres may be a valuable source of variation in breeding programmes, creating novel centromeric haplotypes in blocks of otherwise infrequently recombining allelic variants of genes. Thus, there are circumstances where some centromeric coverage may be advantageous. Introgression from wild species is a routine pre-breeding activity in potato and understanding the structural diversity of pre-breeding material in terms of the extent of introgressed segments is also important. For instance, the recent phased genome assembly of the tetraploid variety C88 (Bao et al. 2022) revealed a high contribution in terms of wild species, including extensive contribution to the centromeric regions. Such linkage drag of centromeric regions whilst introgressing target genes may be useful for diversity or undesirable due to the introgression of non-optimal alleles. Thus, while we do not have enough information to understand the optimal design and utility of centromeric markers for addition to low density marker panels, we can certainly envisage some applications where centromeric coverage might be useful in future versions.

Cost Considerations

As can be seen from Appendix A Table A1, primers, on a per-assay basis, are not actually the highest contributor to cost (although, in our case, they were the largest initial outlay, even when purchased at the minimum synthesis scale). However, the number of loci surveyed does impact both sequencing cost and achievable coverage, and this is the single biggest cost component per assay, whilst sequencing depth contributes to the ability to identify all alleles at a locus. The biggest per-assay cost, when adopting the original GT-Seq protocol as outlined by Campbell et al. (2015), is library normalisation of individual samples subsequent to library construction (Figure 3, step 3A). This process accounts for more than 20% of the per-assay cost. Rather than attempting to find a cheaper approach for this, we normalized the template DNA (in contrast to Campbell et al.'s (2015) procedure), and we tested the protocol on the Extended FRY population both with and without normalization and processed both datasets through SMAP *haplotype-sites*. We obtained similar results with the non-normalized approach, as with the library normalization approach in terms of data and the number of SNPs and haplotypes. There was a more homogeneous coverage among loci but also a higher number of low-output samples (Figure 5b). The efficiency of the library construction was more variable in certain groups of samples than others, and this seemed largely attributable to different DNA extraction runs in different years, as the Extended FRY population was collected over a three-year period. This observation agrees with those of the GT-Seq developers (Campbell et al. 2015), who demonstrated that the cause of this variation was DNA quality among individual samples. Thus, if homogenous high-quality DNA samples can be obtained for an entire experiment, we suggest that normalization could be dispensed with (Figure 3, step 3B). For sets of DNA extracts with variable quality, we recommend following the complete protocol described in this work, including library normalization (Figure 3-step 3A).

Utility of PotatoMASH in Discovery Genetics and Breeding Applications

We empirically tested the effectiveness of the current core set of markers in PotatoMASH by repeating a GWAS analysis on a subset of 279 individuals of the FRY population in which QTLs for the trait fry colour had previously been detected. In that experiment, >40 k ApeK1-derived GBS markers were applied to detect a QTL on chromosome 10 with a smaller additional QTL on chromosome 2. Interestingly, 2279 SNPs identified in the full FRY population generated a similar number of haplotypes (2012) at the 333 core loci when processed with SMAP *haplotype-sites* (non-normalized dataset). We utilised both of these marker sets in GWAS to compare their detection power. In the absence of existing software or models to harness the multiallelic nature of the markers for GWAS in tetraploids, we decided to code the haplotype data in a similar manner to bi-allelic SNPs, with each individual haplotype representing one allele, and the absence of that haplotype representing the other allele. We then performed GWAS, using the same settings with the ~2 k SNPs and the ~2 k haplotypes derived from them. The SNP set detected no QTLs whilst the haplotype set detected the QTL on chromosome 10, but not the QTL on chromosome 2, which was marginal relative to the cut-off threshold in the original study. Aggregating the SNPs into short haplotypes clearly increased the ability of the polymorphic marker set to better describe the real underlying haplotypic structure in the population. This increase in resolution was such that the varietal origin of the haplotypes detecting the QTL could be identified as the variety Valor, which contributed a haplotype that negatively impacted fry colour, something that was not apparent in the original analysis based on the 40 k GBS markers. PotatoMASH is designed to be expandable, by virtue of reconstituting the primer pool with additional markers from experiment to experiment (our original primer stock will support over 85,000 samples to genotype). For instance, it would be possible to increase the marker density of the platform by adding sets of validated marker loci from other studies; e.g., Vos et al. (2015) highlight the fact that the 3763 SNPs they included in the SolSTW 20 k array from the original SolCAP 12 k array have now been shown to work across a wide range of samples, exhibiting low levels of ascertainment bias.

Whilst numerous applications in genetics and breeding require genome-scanning capability, others require the ability to target the presence of specific variants, e.g., to diagnose the presence of specific alleles in MAS. To demonstrate the expandability of the platform and the ability to target additional loci of specific interest, we designed primer sets to target diagnostic polymorphisms for disease and pest resistance of interest to the Teagasc/IPM Potato Group breeding programme and then added these to the core set for application to the Extended FRY population. Amongst 10 target loci (Table 2), the amplification and detection of the target SNP directly from the SNP-calling pipeline was consistently successful. However, three of the ten SNPs did not yield an associated haplotype when processed in SMAP *haplotype-sites*, indicating that such targeted markers (low-frequency SNPs in these three cases) might best be analysed at the SNP level prior to the SNP filtering step to maximise their detection. The additional haplotypes (apart from the one containing the target) can contribute with additional information from a genome-wide scanning context.

We reasoned that core markers should have some ability to detect the presence of haplotypes associated with resistance. To test this, we searched for core markers that yielded similar (>90%) segregation patterns to those markers specifically designed to the R-loci, which would indicate that they are detecting the presence of an introgressed segment carrying the original marker and resistance gene. At the time of writing, a total of 21 loci have been analysed in the Extended FRY population, including the 10 described in this study, as well as further unpublishable proprietary markers for which we could not show results. In total, 11 (52%) of these were found to have a linked core haplotype exhibiting at least 90% of concordance in dosage calls with the target SNP. Thus, given a single discrete target (SNP), a core haplotype with similar information content could be detected in 50% of the cases. Presumably, in instances where the phenotype caused by the underlying gene was either qualitative or a large effect QTL at these loci, it is likely that PotatoMASH would have sufficient power to detect its presence.

Potential Applications for PotatoMASH in Potato Breeding

In this manuscript, we have described PotatoMASH in terms of its ability to efficiently scan allelic variation in the potato genome in a cost-effective manner and explored its potential for GWAS, genetic mapping and diagnostic marker detection. Another potential application that drove us to develop the platform is the need to address cost as a limiting feature in applying genomic prediction to potato breeding. We have previously demonstrated moderate to good levels of predictive ability for the fry colour trait using rrBLUP in the FRY population using the 40 k GBS-derived SNPs mentioned above. However, we also showed that, for this trait, it is possible to identify a small subset of SNPs for processing characteristics that can give moderate predictive ability, albeit lower than that achieved with genome-wide markers (Byrne et al. 2020). The concept of using smaller numbers of markers for prediction in potato has been explored by others. For example, it has been recently shown that “pruning” a larger set of SNPs based on the distinct LD signatures in the population they were applied to could reduce the number to 1500–5000 individual SNPs without loss of information for GWAS and GS in that population (Selga et al. 2021b). Interestingly, the harshest pruning took place close to the centromeres, in line with our strategy of not placing markers in this region. Thus, smaller (and cheaper) marker sets can be utilized for genomic prediction in potato. A recent study in wheat showed that multiallelic haplotypes can improve the accuracy of genomic prediction over single SNPs (Sallam et al. 2020), and separately, it has also been shown that allele dosage information can improve predictive abilities in comparison to using diploidized markers in polyploids (Batista et al. 2022). PotatoMASH combines low cost of application, good marker coverage of the euchromatic portion of the genome, highly discriminatory multiallelic haplotypes, and tetraploid/diploid dosage information at low cost. We are currently exploring how to exploit the aforementioned advances in genomic prediction with these features of PotatoMASH for low-cost genomic prediction in potato breeding. The ability to add targeted markers, as illustrated here, means it could potentially be used for simultaneous genomic and marker-

assisted selection strategies, improving the efficiency of selection in potato breeding (Slater et al. 2014).

Conclusions

PotatoMASH (Potato **M**ulti-**A**llele **S**canning **H**aplotags) efficiently surveys genetic variation throughout the potato genome. It can simultaneously diagnose the presence of target pest resistance markers and track haplotype variation for use in breeding and genetics applications where whole-genome scanning capability is needed at low cost for hundreds to a few thousands of samples.

Data Availability Statement

Fastq files are available in the BioProject database under BioProject ID PRJNA858449.

Acknowledgments

We acknowledge the full support of the Teagasc Potato Breeding Program.

Author Contributions

Conceptualization, D.M. and S.B.; methodology, M.d.l.O.L.-P.; software, T.R.; validation, M.d.l.O.L.-P., C.R.C., L.V., and J.K.; formal analysis, M.d.l.O.L.-P.; investigation, M.d.l.O.L.P.; resources, F.M., S.B., and D.G.; data curation, M.d.l.O.L.-P., L.V., and C.R.C.; writing—original draft preparation, M.O.L.P.; writing—review and editing, S.B. and D.M; visualization, M.d.l.O.L.-P., L.V., and C.R.C.; supervision, D.M.; project administration, M.d.l.O.L.-P. and D.M.; funding acquisition, M.d.l.O.L.-P. and D.M. All authors have read and agreed to the published version of the manuscript.

Supplementary Materials

The following are available online at https://www.mdpi.com/article/10.3390/agronomy12102461/s1.

File S1: PotatoMASH primers.

File S2: PotatoMASH bed file, File S3: Tetraploid_population_haplotypes_discrete_calls_filtered.tsv.

File S4: GRM_GWAS_input.xlsx.

File S5: GWAS.

File S6: Diploid_mapping_population _705_haplotypes_discrete_calls_filtered.tsv.

Figure S7: Diploid mapping population coverage heatmap and haplotype frequency profile.

Table S8: SNP_alleles_in_haplotypes.xlsx.

Appendix A

Table A1. Supplies and costs of PotatoMASH materials used to genotype 765 commercial potato lines in this work. Available prices in 2019–2020 (VAT excl.). Preps is the number of samples that can be processed or were processed (sequencing) with the amount of material in the pack.

PotatoMASH Step and Supplies	Provider/Code	Pack Price (EUR)	Pack Units	Preps	Sample Cost (EUR)
DNA extraction:					
705 samples by CTAB method + 60 samples by sigma Kit					0.247
CTAB Lysis buffer	*Applichem A4150*	85.6	1 L	1000	
GenElute™ Plant Genomic DNA Miniprep Kit	*Sigma G2N10-70KT*	150	70	70	
2 mL tubes for CTAB method	*Greiner 623201CI*	11.43	1000	500	0.023
Isopropanol for CTAB method	*Fisher Chem. P749015*	5.3	1 L	2000	0.003
Ethanol for CTAB method	*Sigma 24105-M*	8.4	2.5 L	3000	0.003
1.5 mL tubes for CTAB method	*Sarstedt 72.690.001*	35	5000	5000	0.007
1 mL tips	*Fisherbrand 11548442*	8.92	1000	916	0.010
200 µL tips	*Sarstedt 70.760.002*	40	10,000	10,000	0.004
96-well PCR plate	*Thermo Sci. 10425733*	58.27	25	2400	0.024
plate adhesive lid	*Greiner 676001*	14.3	100	9600	0.001
Template DNA normalization:					
Quant-iT™ PicoGreen® dsDNA Assay Kit	*Thermofisher P7589*	510	2000	1916	0.266
Nunc™ F96 MicroWell™ Black Plates	*Thermofisher 236105*	112	50	4400	0.025
10 µL tips for quantitation	*Greiner 771290*	9	1000	1000	0.009
200 µL tips for quantitation	*Sarstedt 70.760.002*	40	10,000	10,000	0.004
10 µL filter tips for normalization	*Sarstedt 70.1130.210*	70	1920	1920	0.036
96-well PCR plate	*Thermo Sci. 10425733*	58.27	25	2400	0.024
plate adhesive lid	*Greiner 676001*	14.3	100	9600	0.001
PotatoMASH PCR1:					
QIAGEN Multiplex PCR Plus Kit (100)	*Qiagen 6152*	185	2.55 mL	700	0.264
PotatoMASH Primers (n = 347, 750 uM each)	*IDT*	5119	20 µL	85,714	0.060
96-well PCR plate	*Sarstedt 72.1978.202*	69.75	25	2400	0.029
10 µL filter tips	*Sarstedt 70.1130.210*	70	1920	1920	0.036
Adhesive aluminium foil plate lid	*Sarstedt 95.1995*	50.5	100	9600	0.005
PotatoMASH PCR2:					
100 µL filter tips to dilute PCR1	*Sarstedt 70.760.212*	70.8	1920	1,920	0.037
QIAGEN Multiplex PCR Plus Kit (100)	*Qiagen 6152*	185	2.55 mL	490	0.378
i5 and i7 Primers (n = 96 + 8) at 100 µM	*IDT*	667	300 µL	2875	0.232
96-well PCR plate	*Sarstedt 72.1978.202*	69.75	25	2400	0.029
10 µL filter tips	*Sarstedt 70.1130.210*	70	1920	960	0.073
Adhesive aluminium foil plate lid	*Sarstedt 95.1995*	50.5	100	9600	0.005
Library normalization, **Pooling wells**, **and Concentration-purification:**					
SequalPrep™ Normalization Plate Kit	*Invitrogen A1051001*	1050	10	960	1.094
10 µL tips for binding step	*Greiner 771290*	9	1000	500	0.018
200 µL tips for washing step	*Sarstedt 70.760.002*	40	10,000	10,000	0.004
200 µL tips for elution step	*Sarstedt 70.760.002*	40	10,000	10,000	0.004

PotatoMASH Step and Supplies	Provider/Code	Pack Price (EUR)	Pack Units	Preps	Sample Cost (EUR)
plate adhesive lid	*Greiner 676001*	14.3	100	9600	0.001
Size selection, **Purification**, **Quantification**, **and Pooling plates:**					
Buffer PB	*Qiagen 19066*	87	500 mL	6400	0.014
15 mL tubes	*Sarstedt 62.554.002*	55	500	48,000	0.001
QIAquick PCR purification Kit	*Qiagen 28704*	104.14	50	4800	0.022
1 mL tips	*Fisherbrand 11548442*	8.92	1000	6857	0.001
Wizard SV Gel and PCR Clean-Up System	*Promega A9281*	94	50	4800	0.020
AMPure XP magnetic beads	*Beckman C. A63881*	1326	60 mL	57,600	0.023
100 µL filter tips	*Sarstedt 70.760.212*	70.8	1920	30,720	0.002
Qubit™ dsDNA BR Assay Kit	*Thermofisher Q32853*	275	500	48,000	0.006
Qubit™ assay tubes	*Thermofisher Q32856*	70	500	48,000	0.001
Library quality assessment and Sequencing:					
One Lane Illumina HiseqX 50% PhiX, paired-end 2 × 150 nt reads.	Novogene (UK)	1307	1 lane	765	1.708
Minor inherent expenses:					0.125
Other supplies which individual cost per sample is too low such as gloves, RNAse (macherey 740505, 0.000016 EUR/sample), ddH_2O, one 10 mL pipette to dispense PB buffer, Agarose, TBE buffer, GelRed dye, 100 bp DNA ladder, one scalpel to cut gel slices, 200 uL tips and ethanol to wash the ampure beads, 10 uL tips for Qubit quantitation to pool the final library, and library shipment to UK with coolers.					
Total cost per sample:					**4.882**
Without library normalization:					3.759

Chapter 3

QTL discovery for agronomic and quality traits in diploid potato clones using PotatoMASH amplicon sequencing.

Authors

Lea Vexler[1,2,3]*, Maria de la O Leyva-Perez[1], Agnieszka Konkolewska[1], Corentin R. Clot[2,3], Stephen Byrne[1], Denis Griffin[1], Tom Ruttink[4,5], Ronald C. B. Hutten[2], Christel Engelen[2], Richard G.F. Visser[2], Vanessa Prigge[6], Silke Wagener[6], Gisele Lairy-Joly[7], Jan-David Driesprong[8], Ea Høegh Riis Sundmark[9], A. Nico O. Rookmaker[10], Herman J. van Eck[2,3]*, Dan Milbourne[1]

Affiliations

[1]Teagasc, Crop Science Department, Oak Park, R93 XE12 Carlow, Ireland

[2]Plant Breeding, Wageningen University & Research, P.O. Box 386, 6700 AJ Wageningen, The Netherlands

[3]The Graduate School Experimental Plant Sciences, Droevendaalsesteeg 1, 6708 PB Wageningen The Netherlands

[4]Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Plant Sciences Unit, Caritasstraat 39, 9090, Melle, Belgium

[5]Department of Plant Biotechnology and Bioinformatics, Faculty of Sciences, Ghent University, Technologiepark 71, 9052 Ghent, Belgium

[6]SaKa Pflanzenzucht GmbH & Co. KG, Eichenallee 9, 24340 Windeby, Germany

[7]Germicopa Breeding, 1 Allée Loeiz, 29000 Quimper, France

[8]Meijer Potato, Bathseweg 47, 4411 RK Rilland, The Netherlands

[9]Danespo A/S, Dyrskuevej 15, DK-7323 Give, Denmark

[10]AVERIS Seeds, Valtherblokken zuid 40, 7876 TC Valthermond, The Netherlands

Published in G3: Genes|Genomes|Geneics (2024)

https://doi.org/10.1093/g3journal/jkae164

Abstract

We genotyped a population of 618 diploid potato clones derived from six independent potato-breeding programmes from NW-Europe. The diploids were phenotyped for 23 traits, using standardised protocols and common check varieties, enabling us to derive whole population estimators for most traits. We subsequently performed a Genome-Wide Association Study (GWAS) to identify quantitative trait loci (QTL) for all traits with SNPs and short-read haplotypes derived from read-backed phasing. In this study, we used a marker platform called PotatoMASH (Potato Multi-Allele Scanning Haplotags); a pooled multiplex amplicon sequencing based approach. Through this method, neighbouring SNPs within an amplicon can be combined to generate multi-allelic short-read haplotypes (haplotags) that capture recombination history between the constituent SNPs, and reflect the allelic diversity of a given locus in a different way than single bi-allelic SNPs. We found a total of 37 unique QTL across both marker types. A core of 10 QTL were detected with SNPs as well as with haplotags. Haplotags allowed to detect an additional 14 QTL not found based on the SNP set. Conversely, the bi-allelic SNP set also found 13 QTL not detectable using the haplotag set. We conclude that both marker types should routinely be used in parallel to maximize the QTL detection power. We report 19 novel QTL for nine traits: Skin Smoothness, Sprout Dormancy, Total Tuber Number, Tuber Length, Yield, Chipping Colour, After-cooking Blackening, Cooking Type and Eye depth.

Introduction

Potato is an important food crop and is a key element in the global food security, as well as being a valuable cash crop (FAO Crops statistics database: http://faostat.fao.org/). Given the importance of potato, and the potential impact of factors such as climate change and world population increase, the ability to rapidly and precisely breed potato varieties combining large numbers of favourable traits has been widely recognized. Outbreeding and tetraploidy of modern cultivated potato are complicating factors to achieve greater genetic gains in potato breeding. The potential to rapidly harness recurrent selection to fix favourable alleles and purge deleterious ones across cycles of selection is limited. In response, several groups have started programmes to increase the effectiveness of recurrent selection by breeding at the diploid level (Song and Endelman 2023; Lindhout et al. 2011; Zhang et al. 2021; Bradshaw 2022) through utilisation of self-compatible diploids. Selfing allows fixation of alleles linked to important traits, after which inbreeding depression is addressed by crossing divergent, high performing inbred lines, producing uniform F1 progeny exhibiting hybrid vigour (Zhang et al. 2021; Lindhout et al. 2011; Jansky et al. 2016). Diploid potato has a genetically encoded gametophytic self-incompatibility system (Hosaka and Hanneman 1998a; Kao and McCubbin 1996). The ability to self-fertilize and backcross lines efficiently is mediated by the *Sli* locus, originally described by Hosaka and Hanneman (1998b), and more recently mapped by Clot et al. (2020). The latter study found that *Sli* is not only available in clones derived from *Solanum chacoense*, but also in material derived from the early variety Rough Purple Chili. Hence, the *Sli* gene is widely present in tetraploid varieties and diploid material derived from these varieties.

The “precision breeding” approach exemplified by utilizing self-compatibility to accumulate and fix traits in potato requires tools to manage the genetic diversity at important loci like *Sli* into diploid breeding material. As well as characterizing the genetic location and origin of *Sli*, Clot et al. (2020) developed and validated diagnostic KASP markers to enable efficient marker assisted selection (MAS) for the *Sli* locus. Other traits important to this breeding approach, are those related to sexual polyploidizations (Clot et al. 2023; Clot et al. 2024), as well as tolerance to inbreeding depression (van Lieshout et al. 2020; Zhang et al. 2022b). These resources will facilitate the reproductive aspects of breeding potato at the diploid level, enabling MAS strategies to manage the introgression of key alleles to facilitate the process. In addition to this, it would be useful to develop a resource for genome-based breeding methods to support the improvement of other important traits in potato at the diploid level such as disease resistance, and agronomic and quality traits to develop specific ideotypes to serve different market segments (e.g., fresh consumption, processing, starch). Identifying marker-trait associations is essential to drive MAS for rapid breeding. One powerful strategy to discover markers linked to complex traits is to perform Genome-Wide Association Studies (GWAS).In the last decades, numerous association studies have been conducted on potatoes, mostly at the tetraploid level, reflecting the above desire to characterize important traits directly in breeding-relevant material (Baldwin et al. 2011;

Byrne et al. 2020; D'hoop et al. 2014; D'hoop et al. 2008; Zhang et al. 2022a; Urbany et al. 2011; Li et al. 2010; Lindhout et al. 2011; Malosetti et al. 2007; Prodhomme et al. 2020; Klaassen et al. 2019; Vos et al. 2022; Rosyara et al. 2016; Schönhals et al. 2017; Sharma et al. 2018). Genetic studies in diploid potato have largely been based on mapping specific traits using biparental crosses, with relatively few, and generally smaller scale GWAS studies in diploid germplasm sets (Díaz et al. 2021; Parra-Galindo et al. 2021; Yang et al. 2021), and, to our knowledge, this study is the most extensive diploid potato panel, multi-trait GWAS so far.

In potato, Genotyping-by-Sequencing (GBS) as marker discovery and screening strategy usually yields tens to hundreds of thousands of SNPs: e.g. 186k (Sverrisdóttir et al. 2017), 40k (Byrne et al. 2020), 22.5k markers (Wang et al. 2021); and SNP arrays in potato typically contain up to tens of thousands of markers: e.g., 20k (Vos et al. 2022) and 8.3k markers (Mosquera et al. 2016; Rosyara et al. 2016). In a previous study, we developed a marker system called PotatoMASH (Leyva-Pérez et al. 2022), with the specific ambition of exploring the potential of low cost, genome-wide genotyping for application in potato breeding and genetics. PotatoMASH surveys 339 loci using a multiplex amplicon sequencing approach followed by deep NGS sequencing (2x150bp Illumina sequencing). The question of what is the minimum number of loci that would provide reasonable genome coverage for effective downstream analysis such as GWAS is in the basis of the development of PotatoMASH. It was previously found that "useful" levels of linkage disequilibrium (LD) extended between 0.6 and 1.5 Mb depending on the population under examination and the LD criterion used. In addition, almost no LD decay was observed across the pericentromeric heterochromatin (Vos et al. 2015; Sharma et al. 2018). This is why PotatoMASH was designed to detect variation at 339 loci evenly distributed every 1 Mb across the euchromatic portion of the genome (Leyva-Pérez et al. 2022), so no site could be more than 0.5 Mb from at least one locus. On the other hand, SNPs are almost entirely bi-allelic, and surveying a single SNP locus per megabase will not efficiently survey the diversity of real haplotypes at any one locus. Because of the high SNP density in potato germplasm, PotatoMASH actually yields >2000 SNPs, and additional tools can be used for read-backed phasing (Schaumont et al. 2022), to create short haplotypes (165-180bp) that can be used as a multi-allelic marker system. These multi-allelic haplotags better represent the real allelic composition at a locus and may have better discriminatory power than SNPs for quantitative trait loci (QTL) detection in genome-wide association analysis. Proof of concept of the detection power of PotatMASH was provided by detecting the same QTL associated with fry colour that was originally detected in a GWAS involving 40K GBS-derived SNP markers (Byrne et al. 2020). In addition, we observed that the multiallelic haplotags potentially had better discriminatory power than SNPs in GWAS, since the QTL was only detected when using multiallelic haplotags and not SNPs (Leyva-Pérez et al. 2022).

In this study, we describe a set of 618 diploid potato genotypes, assembled by a consortium of six breeding programmes (DIFFUGAT project https://diffugat.eu/). This material will form the basis of the diploid breeding approaches described above. Phenotypic data were

collected on 23 traits over three years (2019–2021) This collaborative project aims to improve commercially relevant traits in a diploid genetic background with several essential reproductive traits such as (1) self-compatibility to allow fixation of genetic gains, (2) 2n gametes to allow sexual polyploidization and hybridization with varieties, and (3) a high level of male and female fertility.

The objectives of this study were: 1) to characterize this panel for a set of traits that are routinely phenotyped during the selection process in these breeding programmes; 2) to map loci underlying the control of these traits using GWAS; 3) to test haplotags based QTL detection in a wide variety of traits. A longer-term goal is to utilize this information to develop marker-based tools to facilitate selection in this germplasm and extended sets of breeding clones related to it within individual programmes.

Materials and Methods

Plant materials and phenotypic evaluation

We used a panel of 618 diploid potato clones provided by a consortium composed of commercial breeders and research institutes. The panel represents clones from diploid breeding programs, where commercially relevant traits are combined with traits important for diploid breeding, such as fertility, self-compatibility and 2n gamete production. Contributions were made by Meijer Potato, The Netherlands – 225 individuals; Wageningen University, The Netherlands – 134; Danespo A/S, Denmark – 101; SaKa Pflanzenzucht GmbH & Co. KG, Germany – 93; Germicopa Breeding, France – 60; and Averis Seeds B.V., The Netherlands – 17 individuals. Because of the commercial nature of the material, pedigree information could not always be provided. In general, the panel is composed of elite diploid breeding clones, primary dihaploids extracted from tetraploid varieties and donors of resistance and fertility traits.

For intellectual property reasons, the breeding material was not shared between companies. Instead, consortium members evaluated their own material using an augmented design, with replicated checks shared across the six locations over three years (2019-2021). Each company implemented field trial design according to their own system, but the check varieties were included across programmes: Two control varieties were used in 2019 (Lady Claire and Fontane), and two additional control varieties were introduced to the experiments in years 2020 and 2021 (Darling and Laperla). Those four controls were used to estimate the environmental variance across the sites. Some additional controls were introduced locally within each company in accordance with their local protocols for field trials. The size of each experimental unit was 8 plants per plot, with the exception of Averis, who planted 14 plants per plot, and we accounted for this in measurements that are influenced by number of plants: Yield, Total Tuber Number and Dry Matter Content, by rescaling the measurement proportionally to 8 plants. All companies used a planting distance of 75 cm between the ridges and 30 to 35 cm between plants. Fungicide treatment, fertiliser and irrigation was

applied according to each company's own growing protocol and according to needs each season. More experimental information is provided in Suppl. File 1.

An overview of all 23 morphological, agronomic and quality traits examined in this study is shown in Table 1. All consortium members used an agreed standardized set of protocols for scoring each trait. While most phenotyping efforts need no further clarification, some observation methods are briefly outlined below. Tuber length (TPM) was measured by counting how many randomly picked tubers are required to fill a PVC gutter of 1 meter length (Suppl. File 2). This means that higher scores are given to shorter tubers. Enzymatic Browning (EnzB) was scored on strings of tuber tissue 2 hours after being scraped from peeled raw potatoes using a coarse kitchen grater. Presentability of Tubers (PTY) is a holistic trait as defined by breeders' experience and includes regularity and goodness of shape, size, eyes, and skin phenotypes. Skin Smoothness (SkinS) relates to the feel and washability of tubers. Skin Brightness (Gloss) is a visual assessment referring to a glossy or shiny skin finish. Cooking Type (CT) was evaluated by boiling samples of two tubers per plot for 25 minutes. After-cooking blackening (ACB) was assessed on the cooled-down potato one day after cooking. Processing quality was assessed using Chipping Colour data from three treatments: 1. tubers stored at 8°C for 4 months before crisping (QDC1-8), tubers stored at 8°C for 6 months before crisping (QDC2-8) and tubers stored at 4°C for 6 months before crisping (QDC2-4). The colour is assessed for three potato tubers, cut into slices of 1 mm and fried at 180°C until water ("bubbles") has disappeared from the crisps.

Table 1. Overview of phenotypic traits, scales and numbers of genotypes (including controls) that were assessed for each trait and number of observations over three years.

Trait	Abbr.	Scale	Number of genotypes (including controls)	Number of observations over 3 years
Yield	YLD	In kg per plant, fresh weight at harvest	567	1650
Canopy stage 1 6 weeks after planting	Can1	1 = plants have not yet emerged to 9 = largest canopy in the trial	475	1347
Canopy stage 2 10 weeks after planting	Can2	1 = plants have not yet emerged to 9 = largest canopy in the trial	550	1295
Tuber Length	TPM	Tubers per meter count was used with correction table	307	907
Total Tuber Number	TTN	Count of tubers	523	1452
Tuber Shape	TSH	1 = very round, 2 = round, 3 = round-oval, 4 = round-oval to oval, 5 = oval, 6 = oval to long-oval, 7 = long-oval, 8 = long, 9 = very long	569	1646
Yellow Skin Colour	YSC	1 = white, 2 = cream, 3 = light yellow, 4 = yellow, 5 = dark yellow, 6 = brown	536	1542
Yellow Flesh Colour	FC	1 = clear white, 2 = white, 3 = cream, 4 = light yellow, 5 = yellow, 6 = dark yellow, 7 = very dark yellow	549	1575
Eye Depth	EYE	1 = very deep to 9 = very shallow	567	1648
Presentability of Tubers	PTY	1 = very bad to 9 = very good	565	1644
Skin Smoothness	SkinS	1 = rough to 9 = very smooth	567	1488
Skin Brightness	Gloss	1 = dull to 9 = clear	554	1411
Sensitivity to Common Scab	Scab	1 = heavy symptoms to 9 = no symptoms	424	1278
Enzymatic Browning	EnzB	1 = ink black, 2 = uniformly black, 3 = discolouration to black, 4 = darkening of red and grey discolouration, 5 = bright red and dark grey discolouration, 6 = start of red/grey discolouration, 7 = clear start of discolouration, 8 = very slight discolouration, 9 = no discolouration	552	1467
Cooking Type	CT	2 = very floury, loose boiling, sloughing, 4 = floury, crumbly and fairly loose, 6 = slightly floury and fairly firm, 8 = not floury, firm cooking, 9 = extreme firmness	555	1393
After-cooking blackening	ACB	1 = very dark to 9 = pure colour (no darkening at all)	558	1498
Chipping Colour 1st time point stored at 8°C	QDC_1-8	1 = very dark to 9 = pure colour (no darkening at all)	559	1593
Chipping Colour 2nd time point stored at 8°C	QDC_2-8	1 = very dark to 9 = pure colour (no darkening at all)	367	857
Chipping Colour 2nd time point stored at 4°C	QDC_2-4	1 = very dark to 9 = pure colour (no darkening at all)	505	1353
Dry Matter Content	DM	% relative to fresh weight	566	1626
Sprout Dormancy	SD	1 = heavy sprouting (early) to 9 = no sprouting	536	1416
Tuber Regularity	REG	1 = bad to 9 = good	565	1646
Maturity	MAT	1 = plants still green and flowering to 9 = plants reached senescence	427	857

Statistical analysis of phenotypic data

All statistical analyses and data visualizations were performed using R version 4.2.1 unless otherwise specified in results. Visual inspection of the distribution of the data and quantile–quantile (QQ) plots of residuals versus quantiles revealed some obvious deviations from homoscedasticity or normality in the continuous traits: Yield, Tuber Length and Total tuber number per plant (Figure 1). The data of those traits were transformed with Yeo-Johnson transformation using the R package “car” (Fox and Weisberg 2019). Although the majority of the traits were measured on an ordinal scale, an inspection of diagnostic plots for residuals indicated no strong violations of the assumption of normal error distributions and were all analyzed as quantitative traits, as previously performed (D’hoop et al. 2008), assuming the error variation to be normally distributed with constant variance.

Check varieties were used in the estimates of the Best Linear Unbiased Estimators (BLUEs) of phenotypic means of all 23 traits across years and locations but were excluded from the GWAS. We used a multiple linear regression package lme4, using the lm function (Bates et al. 2015) to calculate the BLUEs with the following equation:

Trait= Genotype + Year + Location + Location*Year + error

Where genotype is the clone name and does not include any genetic information such as pedigree due to intellectual property rights and Location is the site of each company. Location*Year effect was applied when analysing data from more than one company. All independent variables: Genotype, Year, Location and Location*Year, were considered as fixed effects due to the low number of levels.

Least square means, calculated for the BLUEs with the R package “emmeans” (Lenth et al. 2021), served as the final phenotypic data used in the association analysis.

A Pearsons correlation matrix between the vegetation indices and the vegetative growth parameters was generated using the package corrplot for R. Correlation coefficients were tested at $p = 0.05$.

Broad-sense heritabilities (H^2) were calculated for each breeding population separately on an entry-mean basis according to the formula:

$$H^2=\sigma_g^2/(\sigma_g^2+\sigma_{gy}^2/n_{year}+\sigma_e^2/n_{year}),$$

Where σ_g^2 is the genotypic variance, σ_{gy}^2 is the genotype-by-year variance, σ_e^2 the error variance and n_{year} is the number of years.

The k-matrix of the genomic data was calculated with the R package GWASpoly (Rosyara et al. 2016) and the modelled Least square means were used to calculate the Marker-based heritabilities, (h^2_{SNPs} and $h^2_{haplotags}$) with the R package “heritability” with the *marker_h2_means* function (Kruijer and White 2023; Kruijer et al. 2014).

Genotypic Data

Data collection with PotatoMASH

Leaf material was sampled in 2019, the first year of the field trials, freeze-dried and stored with silica gel until use. Approximately 5 mg of dry tissue was used to extract DNA with Mag-BIND® Plant DNA DS Kit (Omega-VWR M1130-00, Philadelphia, USA), using the KingFisher Flex automated extraction & purification system (Thermo Scientific, Austin, TX, USA).

PotatoMASH libraries were obtained and haplotyping was performed as in Leyva-Pérez et al. (2022) (https://www.protocols.io/view/potatomash-library-construction-e6nvw53zdvmk/v2) with the following adjustments to the bioinformatics pipeline: Merged and filtered reads were mapped to the *S. tuberosum* genome v6.1 (Pham et al. 2020). Variant calling were then filtered with a minimum allele frequency of 0.01 and a maximum of 0.99: *vcftools –bcf PotatoMASH.bcf –out PotatoMASH –min-alleles 2 –max-alleles 2 –recode –recode-INFO-all –minQ 30 –minDP 6 –maf 0.01 –max-maf 0.99 –remove-filtered-all –max-missing 0.5*.

Haplotypes nomenclature is given by the software SMAP (Schaumont et al. 2022) as a binary string code for the set of SNPs called in a specific locus, were the reference allele of each SNP is coded as “0” and the alternative allele is coded as “1” in a specific haplotype (Figure 4c). The final haplotag name is the PotatoMASH locus name plus the binary string in which 0 means same base as reference genome, 1 is alternative base and “-“ is an indel at that SNP position (e.g. C1_1_000110-10).

For the 334 polymorphic loci, the average locus correctness score (number of samples with sum of discrete haplotag dosage calls equals 2, divided by total number of samples with sufficient read depth for that locus, expressed as percentage) was 92. SMAP also calculates the sample correctness score per sample (number of loci where the sum of discrete haplotag dosage calls equals 2, divided by the total number of loci with sufficient read depth, expressed as percentage). Since the average locus correctness score was high for the 334 loci, we assumed that individuals with low sample correctness score would be due to technical errors or putative cross contamination. Therefore, we removed 21 genotypes with a sample correctness lower than 40. A final panel of 558 genotyped individuals were used for the GWAS.

Population structure

Population structure was evaluated using a principal component analysis (PCA) calculated with Plink 1.9 using SNPs with a Minimum Allele Frequency >0.01 (Purcell et al. 2007). The population genetic structure was assessed using the Bayesian clustering method implemented in STRUCTURE version 2.3.4 (Pritchard et al. 2000). An admixture model and correlated allele frequencies were chosen for estimating the proportion of ancestral contribution in each accession. We tested various K-values ranging from 1 to 10 with 3 independent replications at each K, 10,000 generations burn-in period and 10,000 Markov Chain Monte Carlo (MCMC) repetitions. Calculation of Delta K: 1. Mean L(K) (±SD) was

done over three independent runs for each K value 2. Rate of change of the likelihood distribution (mean± SD) was calculated as L′(K) = L(K) – L(K – 1). 3. Absolute values of the second order rate of change of the likelihood distribution (mean ± SD) were calculated according to the formula: |L″(K)| = |L′(K + 1) – L′(K)|. 4. ΔK calculated as ΔK = mean|L″(K)|/ sd[L(K)] (Evanno et al. 2005). Visualizing admixture plot was done with the fastSTRUCTURE software distruct.py function (Raj et al. 2014).

GWAS

Two datasets were used for the GWAS. We first identified SNPs across the sequenced amplicons and used these as a data set in a GWAS. We then used this SNP set to construct short haplotypes with SMAP *haplotype-sites* tool (see section “data collection with PotatoMASH”) combined with discrete genotype calling, that were then used simply as presence-absence markers for GWAS. The distinct haplotags were treated as "pseudoSNPs" for the purpose of the analysis.

Association analysis for both SNP and haplotag data was done with the R package GWASpoly (Rosyara et al. 2016). Population structure was controlled using the K model and QQ plots were used to assess if there was sufficient control of population structure. The function GWASpoly with additive and non-additive models was used to test for association at each marker. Marker curation was carried out using the maximum genotype frequency option with default parameter setting (geno.freq = 1-5/N, where N is the number of genotypes), so markers present in fewer than five individuals are removed. The genome-wide false discovery rate was controlled using the M.eff method (a Bonferroni-type correction but using an effective number of markers that accounts for Linkage Disequilibrium (LD) between markers) at level = 0.05. We did not use the leave-one-chromosome-out (LOCO) approach due to the inflation of the P-values as observed with the QQ plots.

Results

Phenotypic data

Taken together, phenotyping of the panel of 618 diploids resulted in 32,590 data points, collected over three years, across six locations, for a total of 23 agro-morphological and quality traits (Figure 1, Table 2). These data were unbalanced given that some locations/breeders focussed on a single niche market (e.g. starch). A strong year-by-location interaction was observed (Suppl. File 3) using control varieties planted across all sites. From these raw data best linear unbiased estimators (BLUEs) were calculated while taking the year-by-location interaction into account with the regression models.

We could not calculate the broad sense heritability (H^2) across all companies, as only the control varieties were shared. The heritabilities presented in Table 2 are the average of the estimated trait heritability for each company and varied mostly between moderate to high values, ranging from 50 to 90%. Traits largely controlled by single loci such as Tuber shape (TSH), Yellow flesh colour (FC), and Maturity (MAT), typically show H^2 between 82% and 90%, according to expectations. Some of the complex polygenic traits like Dry Matter

Content (DM) and Yield, also show an exceptionally high H^2 of 85-89%. Furthermore, the majority of the processing and quality traits such as Enzymatic Browning (EnzB), Cooking Type (CT), After-cooking Blackening (ACB), and Chipping Colour showed moderately high H^2 values (73-84%). Marker-based heritability, were also calculated using both markers types, SNPs and haplotags, and were lower than the broad sense heritability for all traits (Suppl. File 10)

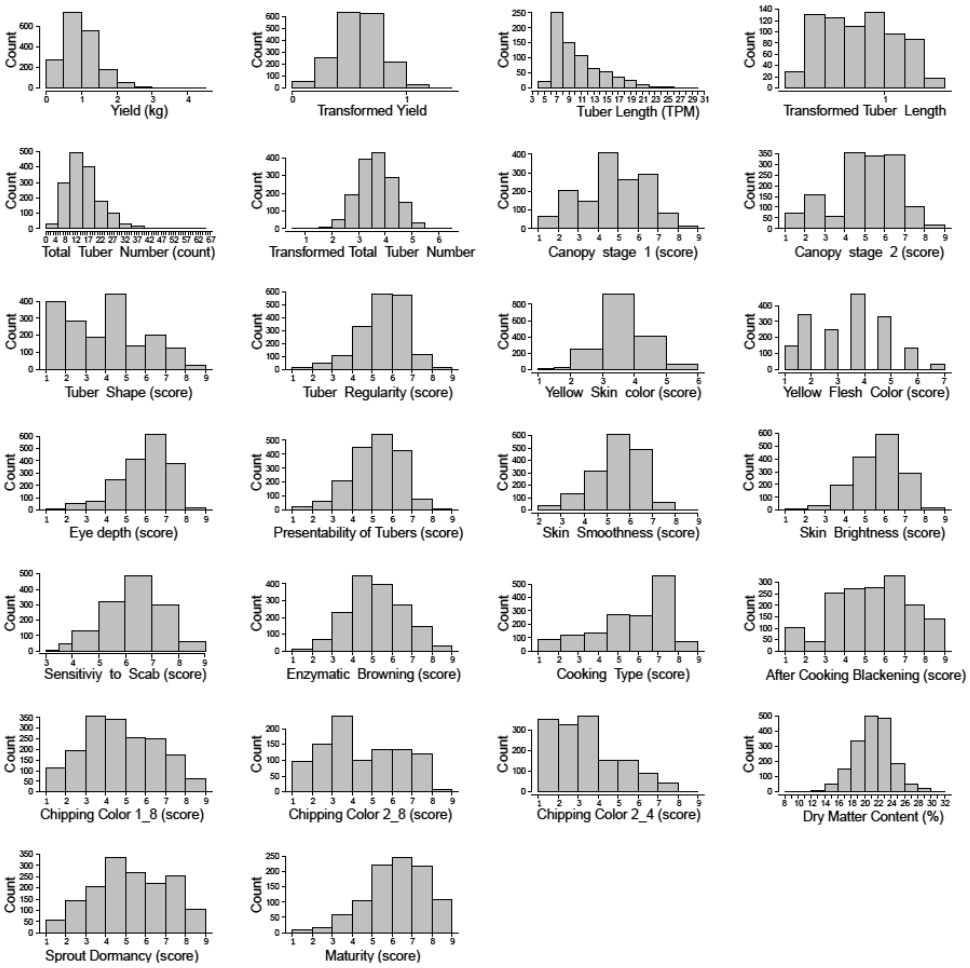


Figure 1. Frequency distribution histograms for all traits of this study, across all companies. The horizontal axis indicates the data range of traits, and the vertical axis indicates the frequency of individuals. For traits Yield, Tuber Length and Total Tuber Number per plant: the grey histograms represent the raw data, and the blue histograms represent the transformed data that we used in the downstream analysis.

Table 2: Mean, standard deviation (SD), minimum (Min), maximum (Max) values and Broad sense (H^2) Heritability (%) for all traits:

Trait	Min	Max	Mean	SD	Average H^2 across all companies	Number of companies tested
Canopy stage 1	1	9	5.24	1.69	60.1	5
Canopy stage 2	1	9	5.53	1.73	75.2	6
Yield	0.003	4.26	0.98	0.5	85.5	6
Tuber Length	5.37	28.11	10.37	4.01	87.1	4
Total Tuber Number	0.75	61.25	15.42	6.72	69.2	5
Tuber Shape	1	9	4.4	2.07	90.4	6
Tuber Regularity	1	9	6.05	1.23	64.9	6
Yellow Skin Colour	1	6	4.12	0.8	63.6	6
Yellow Flesh Colour	1	7	3.6	1.5	88.8	6
Eye depth	1	9	6.47	1.31	82.6	6
Presentability of Tubers	1	9	5.65	1.24	74.5	6
Skin Smoothness	2	9	5.95	1.09	60.3	6
Skin Brightness	1	8	5.58	1.09	62.1	5
Sensitivity to Common Scab	3	9	6.76	1.16	50.6	6
Enzymatic Browning	1.5	9	5.51	1.36	81.7	6
Cooking Type	2	9	6.62	1.68	74.3	6
After-cooking Blackening	1	9	5.9	1.9	72.7	6
Chipping Colour 1_8	1	9	5.25	1.86	77.8	6
Chipping Colour 2_8	1	9	4.93	1.96	83.6	3
Chipping Colour 2_4	1	9	3.82	1.77	81.8	4
Dry Matter Content	8.35	31.08	21.35	2.84	88.7	6
Sprout Dormancy	1	9	5.81	1.82	74.9	5
Maturity	1	9	6.6	1.44	82.0	4

Correlations between Traits

The correlations between traits are shown in Figure 2, and are based on the phenotype estimated means (Suppl. File 5). The highest positive correlations were observed between Skin Smoothness and Skin Brightness (while both had a negative correlation with Yellow Skin Colour). High correlations were also observed for tuber visibility traits such as Tuber Regularity, Tuber Presentability, and Eye depth. Yield showed a negative correlation with Maturity and a positive correlation with Canopy development. Canopy stage 1 and Canopy stage 2 positively correlated with Total Tuber Number.

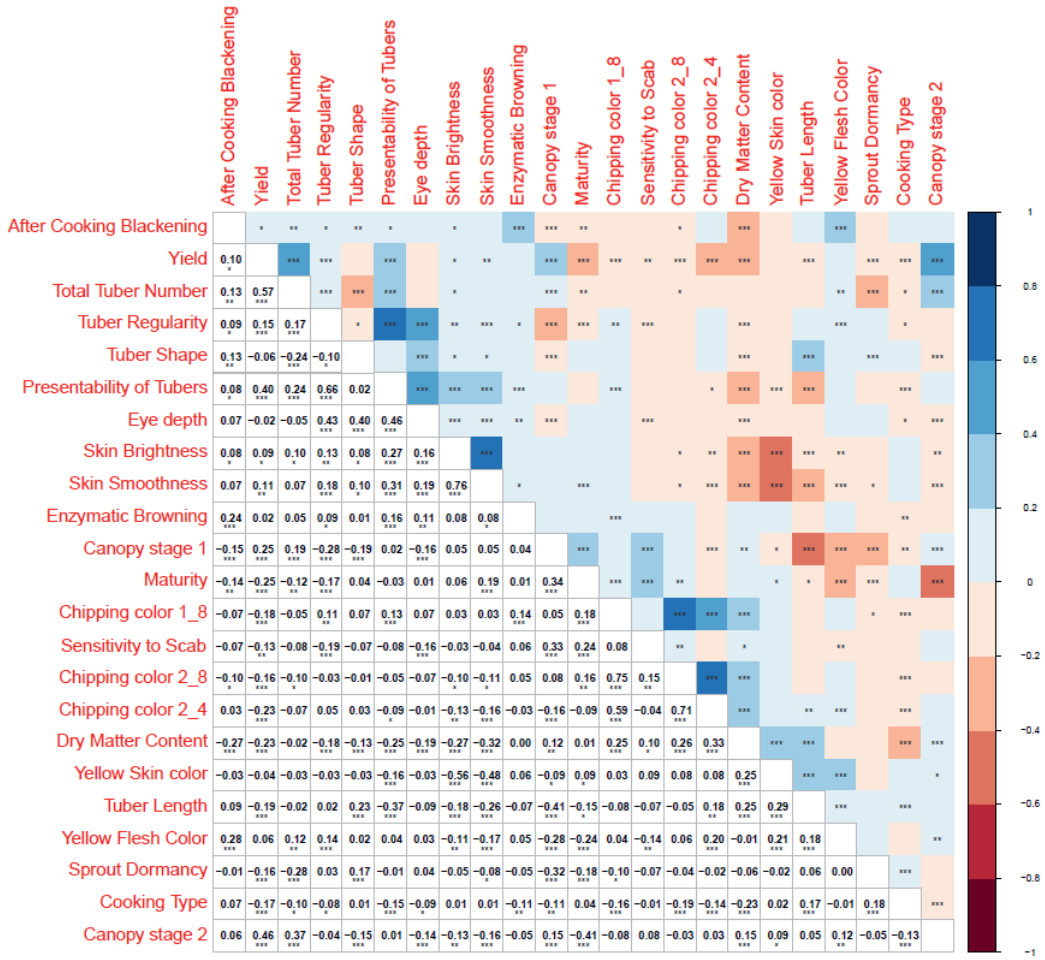


*Figure 2: Matrix of pairwise Pearson's correlation between all traits. Positive correlations are displayed in blue and negative correlations in red. Colour intensity is proportional to the correlation coefficients according to the scale displayed on the right. Marking of significance level: ***0.001, **0.01, *0.05*

Genotyping, variant calling, and genetic diversity

After merging and filtering reads we retained on average 275,855 reads per sample, which corresponds to an average of 813 reads per amplicon per sample, well exceeding the required minimum of 20x read depth recommended for SMAP haplotype calling for diploids. The amplification efficiency of the primer pairs (either low or high) was consistent across most samples (Figure 3).

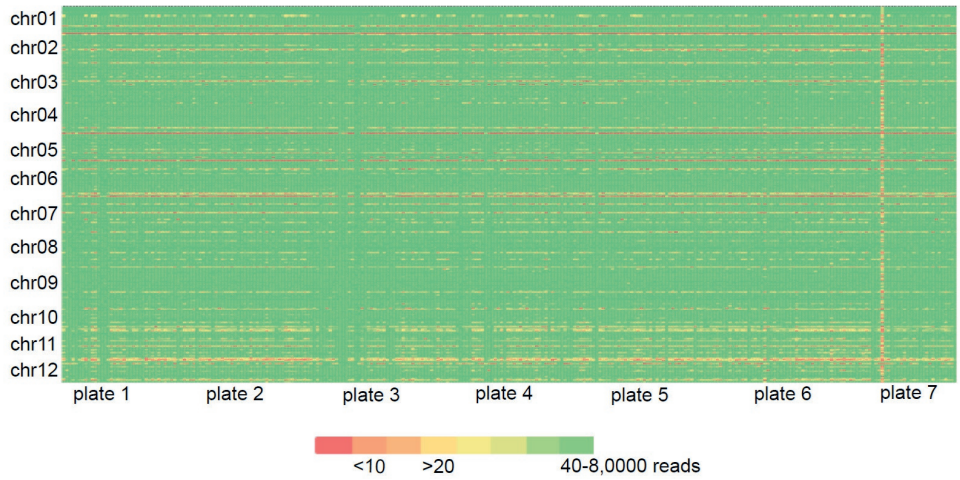


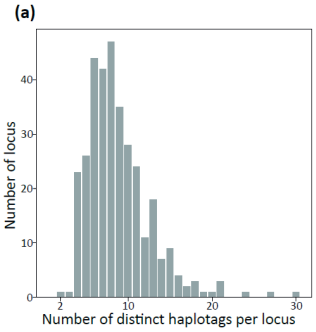
Figure 3. Coverage of the 339 PotatoMASH core loci. Heat map of the number of merged and filtered reads of 618 samples (in columns, each plate is of ~96 samples) that mapped to each locus (in rows).

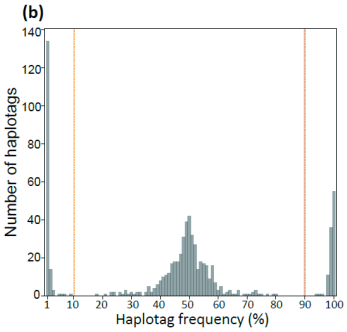
After filtering, 2730 SNPs were identified across the panel. Out of 339 PotatoMASH target loci, SMAP *haplotype-sites* could identify 334 loci with polymorphic, multi-allelic haplotypes. A total of 2955 short multi-allelic haplotags were identified across the panel, ranging from 2–30 haplotags per locus, while most loci had 8-9 haplotags per locus (Table 3, Figure 4a). This is higher than previously reported by Leyva-Pérez et al. (2022) in a tetraploid population where 2-14 haplotags per locus (on average 6 haplotags per locus) were reported. The higher haplotype diversity suggests higher genetic diversity in the used diploid panel.

As expected in our diploid panel, two haplotags (either homozygous or heterozygous) were successfully called at each locus, for each individual, in 91% of cases. SMAP analyses the relative read depth per haplotag per locus per individual, and outputs the distribution across all loci to check, if that fits the typical frequency spectrum expected for diploids (Figure 4b). 39 of the individuals showed a tetraploid typical frequency spectrum and were excluded. As final output, we obtained a table with discrete dosage calls for each haplotag per locus, per sample (Figure 4c), which was used for downstream analysis.

Table 3. Summary of genotyping and variant calling with PotatoMASH.

Total samples	**558**
SNPs called	7503
SNPs filtered	2730
Polymorphic loci	334
Number of haplotypes	2955
Haplotypes per locus	2-30 (8.8
2 haplotags called per locus per	91%





(c)

Locus	Haplotag	Sample_1	Sample_2	Sample_3
Locus 1	00000000	1	2	0
Locus 1	00000010	1	0	1
Locus 1	00000100	0	0	0
Locus 1	00011010	0	0	0
Locus 1	00100000	0	0	0
Locus 1	01000000	0	0	1
Locus 2	000000	0	2	1
Locus 2	000010	0	0	0
Locus 2	000100	1	0	0
Locus 2	010000	0	0	0
Locus 2	011000	1	0	1
Locus 3	000000	2	0	1
Locus 3	001100	0	0	0
Locus 3	001110	0	1	0
Locus 3	010000	0	1	1

Figure 4. (a) Haplotag diversity distribution of 334 loci across the individuals in the panel. (b) Haplotag frequency spectrum of one individual, the haplotag frequency is calculated by the relative read depth (%) for each haplotag within its locus. (c) Example of tabular data generated by SMAP haplotype-sites with 3 genotypes (samples), 3 loci, 15 haplotags, and diploid discrete dosage calls for each locus/sample. Loci 1 and 3 include haplotags not detected in samples 1-3 but in other genotypes not shown (samples 4-558).

Population structure

We examined population structure by Principal Component Analysis (PCA) using the SNP data and observed two main clusters, with separation mainly occurring on the 1st principal component which explains ~17% of the genetic variation, indicating that the diploid population of Meijer deviates from the gene pools of the other breeding programs (Figure 5a). We also examined the underlying population structure of the panel through Bayesian-based approach using STRUCTURE v 2.3.4. and with the log mean probability and deltaK (change in log probability) per K (number of sub-populations) generated the highest peak at K = 2 (Figure 5c), and this confirmed the conclusion of two sub-populations. We therefore decided to perform QTL discovery using three sets of potato genotypes: the “full panel”, the sub-populations "only Meijer" and the rest not including Meijer (referenced as "no Meijer"). In this way, we were hoping to capture QTL that were robust across all subpopulations in addition to subpopulation-specific QTL.

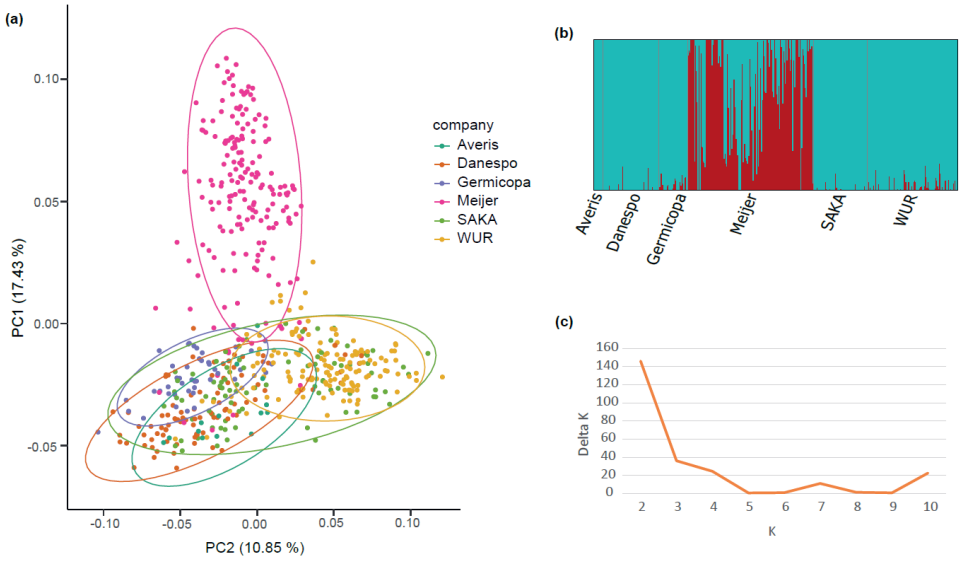


Figure 5. (a) Principle Component Analysis (PCA) with SNP data of all six companies. (b) Estimation of hypothetical sub-populations using K-values. (c) The number of identified sub-populations (K) versus DeltaK estimated based on Evanno method

GWAS of multiple traits

To capture all the potential QTL, we performed six GWAS (three genotype-sets described above with the two marker-sets, SNPs and haplotags).

We identified 37 QTL for 20 out of 23 traits. For three triats: Tuber regularity, Skin brightness and Presentability of tubers we didn’t detect QTL . Of the 37 QTL identified, only 10 QTL were detected with both SNPs and haplotags. 14 QTL were only detected by haplotags, and

13 QTLs were only detected by SNPs (Figure 6; Table 4). The full information of all the significant markers, including their marker's effects, are provided in Suppl. File 7.

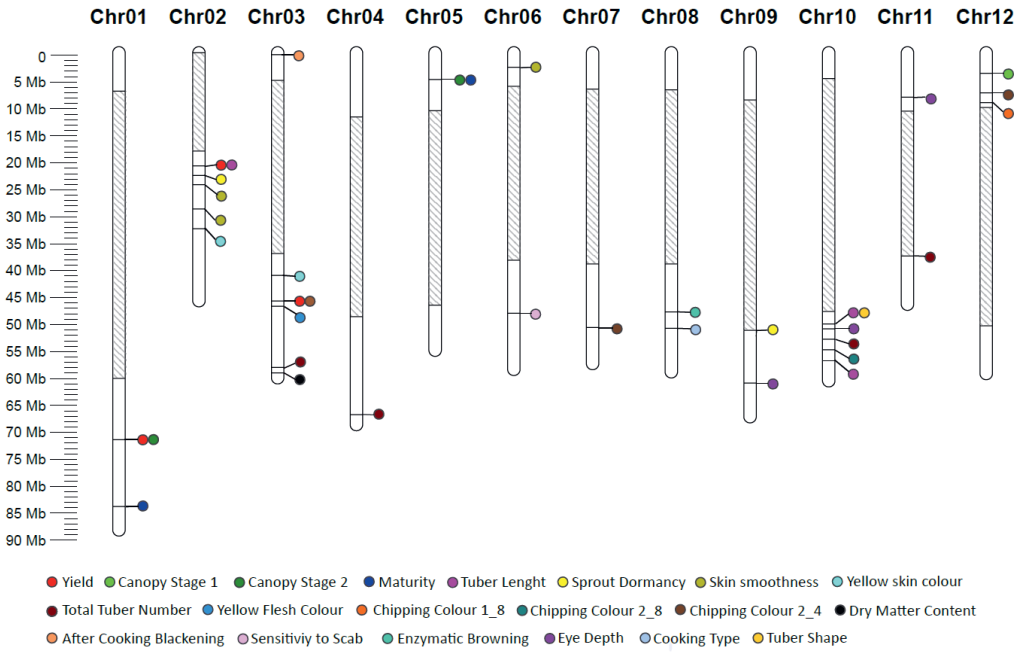


Figure 6. Physical map with the positions (Mb) of all QTL. Grey regions on the chromosome indicate pericentromeric heterochromatin, without PotatoMASH amplicons.

Differences in QTL detected across populations

Differences in the sets of QTL were observed across populations. Only five QTL, for Tuber Shape, Eye depth, Yellow Flesh Colour (two QTL) and Maturity were detected across all three populations with at least one marker shared for each trait across all populations. Another 15 QTL were detected either in the "no Meijer" sub-population (6) or in the "only Meijer" sub-population (9). Eleven QTL were shared between the full panel and one of the other two sub-populations. Ten QTL were detected with the full panel, but were not significant in either sub-population.

Table 4. Overview of the 37 QTLs for the 20 trait, with the significant markers for each population: colours: red – QTL only with SNPs, blue- QTL only with haplotags, black- QTL with both. Columns from left to right; Trait, given name of QTL, QTL location in potato genome DMv6, name of haplotag and SNP for each population, previously reported QTL and positions in Mb when available, in potato genome DMv4 and Literature column citing the previous works in which these QTLswere found.

	QTL		full panel		no Meiljer		only Meijer		Literature	
Trait	Name	chr (Mb in DMv6.1)	Haplotags	SNPs	Haplotags	SNPs	Haplotags	SNPs	chr (Mb in DMv4.3 when available)	Reference
Yield	YLD_C1_19	chr01 (71.77)					C1_19_00110		1, 2, 5, 6, 9, 7, 11, 12	(Bradshaw et al. 2008; da Silva Pereira et al. 2021b; Manrique-Carpintero et al. 2015; McCord et al. 2011; Rak et al. 2017; Schönhals et al. 2017)
	YLD_C2_4	chr02 (20.99)						chr02_20959684 chr02_20959746		
	YLD_C3_17	chr03 (46.06)	C3_17_011000	chr03_46058754	C3_17_011000	chr03_46058754				
Skin Smoothness	SkinS_C2_8	chr02 (24.47)			C2_8_0000001000	chr02_24470953				
	SkinS_C2_13	chr02 (28.96)	C2_13_100010000		C2_13_100010000					
	SkinS_C6_3	chr06 (2.68)	C6_3_0000101-00		C6_3_0000101-00					
Cooking Type	CT_C8_21	chr08 (51.08)					C8_21_0000000000		1, 2, 6, 9, 10, 11	(D'hoop et al. 2014; D'hoop et al. 2008; Kloosterman 2006)
After-cooking Blackening	ACB_C3_1	Chr03 (3.38)		chr03_337574 chr03_337585					1, 2, 3, 4, 5, 6, 7, 11	(Bradshaw et al. 2008; D'hoop et al. 2014; D'hoop et al. 2008)
Dry Matter Content	DM_C3_30	chr03 (59.35)						chr03_59353418	2, 3, 5, 6, 7, 8	(Bradshaw et al. 2008; da Silva Pereira et al. 2021b; McCord et al. 2011)
Canopy Stage 1	Can1_C12_0	chr12 (3.78)		chr12_3783754		chr12_3783754				
Canopy Stage 2	Can2_C1_19	chr01 (71.77)					C1_19_00110			
	Can2_C5_6-C5_8	chr05 (4.94-7.25)	C5_6_0000000	chr05_4941391 chr05_4941406 chr05_7251491 chr05_7251555			C5_6_0000000 C5_7_0001111110	chr05_4941391 chr05_4941406 chr05_6204154 chr05_7251491 chr05_7251555		
Enzymatic Browning	EnzB_C8_18	chr08 (48.05)					C8_18_00100110010010100		1, 3, 4, 5, 6, 7, 8, 11	(D'hoop et al. 2014; D'hoop et al. 2008; Urbany

	QTL		full panel		no Meiljer		only Meijer		Literature	
Trait	Name	chr (Mb in DMv6.1)	Haplotags	SNPs	Haplotags	SNPs	Haplotags	SNPs	chr (Mb in DMv4.3 when available)	Reference
										et al. 2011; Werij et al. 2007)
Tuber Shape	TSH_C10_7-C10_12	chr10 (48-53.13)	C10_7_011110000 10, C10_8_000111110 0, C10_8_000111111 0, C10_8_011000000 0, C10_9_000101000 100000 C10_9_011100000 110110	chr10_49148246 chr10_49148249 chr10_49148293 chr10_49148305 chr10_49148316 chr10_50323192 chr10_50323244	C10_8_0001111100 C10_9_00010100010000 0 C10_12_00010000000	chr10_49148293 chr10_49148305 chr10_49148316 chr10_50323192 chr10_50323244	C10_8_0001111110 C10_8_0110000000 C10_9_00010100010000 0 C10_9_01110000011011 0	chr10_49148246 chr10_49148249 chr10_49148293 chr10_49148305 chr10_49148316 chr10_49148341 chr10_50323151 chr10_50323153 chr10_50323192 chr10_50323225 chr10_50323244 chr10_50323263	10 (48.7)	(Pandey et al. 2022; Rosyara et al. 2016; Sharma et al. 2018; van Eck et al. 1994a)
Eye depth	EYE_C9_21-C9_23	chr09 (61.25-62.91)					C9_21_000100010100 C9_23_000001	chr09_61254396 chr09_62045322	3, (53.1), 5 (43.9), 10 (48.6)	(Li et al. 2005; Pandey et al. 2022; Rosyara et al. 2016; Sharma et al. 2018; ŚLiwka et al. 2008)
	EYE_C10_6-C10_10	chr10 (4.74-51.2)	C10_9_000101000 100000	chr10_4741749 chr10_49148293 chr10_49148305 chr10_49148316 chr10_50323244		chr10_4741749 chr10_49148293 chr10_49148305 chr10_49148316	C10_8_0001111110 C10_9_01110000011011 0 C10_9_00010100010000 0 C10_10_01010000	chr10_49148246 chr10_49148293 chr10_49148305 chr10_49148316 chr10_50323151 chr10_50323153 chr10_50323225 chr10_50323244 chr10_50323263		
	EYE C11_8	chr11 (8.19)		chr11_8190769						
Maturity	MAT_C1_32	Chr01 (84.18)		chr01_84177600					1, 2, 3, 5, 7	(da Silva Pereira et al. 2021b; Kloosterman et al. 2013; McCord et al. 2011)
	MAT_C5_6-C5_7	chr05 (4.94)	C5_6_0000000 C5_6_0011010 C5_7_0001111110	chr05_4941391 chr05_4941406 chr05_4941464	C5_6_0000000	chr05_4941391 chr05_4941406	C5_6_0000000 C5_7_0001111110	chr05_4941391 chr05_4941406 chr05_6204154		
Tuber Length	TPM_C2_4	chr02 (20.96)				chr02_20959691			10 (50), 9	(Zhang et al. 2022a)
	TPM_C10_8-C10_10	chr10 (49.15-51.02)	C10_8_000111111 0 C10_10_00100100				C10_8_0001111110 C10_9_00010100010000 0 C10_9_01110000011011 0 C10_10_01010000	chr10_49148246 chr10_49148249 chr10_49148293 chr10_49148305 chr10_49148316 chr10_49148341 chr10_50323151 chr10_50323153 chr10_50323192 chr10_50323225 chr10_50323244 chr10_50323263		

	QTL		full panel		no Meiljer		only Meijer		Literature	
Trait	Name	chr (Mb in DMv6.1)	Haplotags	SNPs	Haplotags	SNPs	Haplotags	SNPs	chr (Mb in DMv4.3 when available)	Reference
	TPM_C10_16	chr10 (57.02)	C10_16_000000000000							
Total Tuber Number	TTN_C3_29	chr03 (58.36)		chr03_58369063 chr03_58369113		chr03_58369063 chr03_58369113			1, 4, 5, 6, 8, 9, 12	(Zhang et al. 2022a; Manrique-Carpintero et al. 2015)
	TTN_C4_33	chr4 (67.1)	C4_33_00100010		C4_33_00100010					
	TTN_C10_12	chr10 (52.17)	C10_12_00000100000							
	TTN_C11_12	chr11 (37.67)			C11_12_000100000					
Sensitivity to Common Scab	Scab_C6_18	chr06 (48.3)	C6_18_011111111110		C6_18_011111111110				1, 2, 3, 4, 5, 6, 10, 11, 12	(Bradshaw et al. 2008; Braun et al. 2017; da Silva Pereira et al. 2021a; Yuan et al. 2020; Zorrilla et al. 2021)
Chipping Colour 1_8	QDC1-8_C12_11	chr12 (9.26)		chr12_9258876					for 'off the field': 4 (68), 10 (55.2)	(Byrne et al. 2020)
Chipping Colour 2_8	QDC2-8_C10_14	chr10 (55.08)				chr10_55081844			1, 6, 10, 11	(Bradshaw et al. 2008; D'hoop et al. 2014)
Chipping Colour 2_4	QDC2-4_C3_17	chr03 (46.06)			C3_17_000010				11, 6,	(Bradshaw et al. 2008)
	QDC2-4_C7_23	chr07 (50.94)						chr07_50949343 chr07_50949356		
	QDC2-4_C12_9	Chr12 (7.44)	C12_9_0010010001100							
Yellow Flesh Colour	FC_C3_15-C3_20	chr03 (44.04-49.75)	C3_15_00000000 C3_17_000000 C3_18_00000000 C3_18_00011000 C3_20_000000	chr03_44040545 chr03_46058758 chr03_47024967 chr03_47024968 chr03_49750222	C3_15_00000000 C3_17_000000 C3_18_00000000 C3_18_00011000	chr03_45076941 chr03_47024967 chr03_47024968 chr03_49750222	C3_17_000000 C3_18_00000000 C3_18_00011000	chr03_47024967 chr03_47024968	1 (63.8), 3 (44.1), 3 (48.5),3 (49.3), 3 (50.8)	(Pandey et al. 2022; Pham et al. 2020; Sharma et al. 2018)
Yellow Skin Colour	YSC_C2_17	chr02 (32.62)	C2_17_0010010							
	YSC_C3_12-C3_13	chr03 (41.3-42.87)		chr03_41301966 chr03_41302030		chr03_41301966 chr03_41302030 chr03_42866260				
Sprout Dormancy	SD_C2_6	chr02 (22.77)			C2_6_01101	chr02_2276673 chr02_22766770			1, 3, 4, 5, 6, 7, 8, 11	(Bradshaw et al. 2008; D'hoop et al. 2014; D'hoop et al. 2008; Urbany et al. 2011)
	SD_C9_11	chr09 (51.47)		chr09_51478036 chr09_51478037 chr09_51478039						

Identification of previously characterised QTL

Some of the traits evaluated here, were previously described in detail in the literature and enabled us to validate our approach (Figure 7). Indeed, we identified a highly significant QTL ($-\log_{10}(p\text{-value})=14.36$) on chr10 across the three phenotypic datasets for Tuber Shape, which was detected both with SNP and haplotag markers (Figure 7, Suppl. File 7). This QTL corresponds to the well described *Ro* locus (van Eck et al. 1994a; van Eck et al. 2022).

The well-known *Y* (Yellow) locus and the causal gene involved in yellow flesh colour beta-hydroxylase (*Chy2* or *BCH)* (Bonierbale et al. 1988a; Brown et al. 2006; Thorup et al. 2000; Wolters et al. 2010), One isoform (PGSC0003DMG400009501) of the *Bch* gene was reported to be located at 44.1 Mb in DMv4.03 (Pandey et al. 2022) and aligns with position 42.9 Mb on chr03 of the DMv6.1 reference genome sequence (Pham et al. 2020). We identified one QTL for Yellow flesh colour on chr03 spanning the region from 44.04 Mb to 49.75 Mb (PotatoMASH loci C3_15 to C3_20), peaking at 47.03 Mb (PotatoMASH locus C3_18) with LOD score (-log10(p-value)) of 44.66 for the significant marker chr03_47024967 (Suppl. File 7). Two additional PotatoMASH loci on chr03 (C3_11 at 40.41 Mb; and C3_30 at 59.35 Mb) also showed significant associations with Yellow Flesh Colour.

Furthermore, we also detected a QTL for Maturity on chr05, peaking at 4.94 Mb (PotatoMASH locus C5_6). The haplotag C5_6_0011010 and the three SNPs of this haplotag (chr05_4941391, chr05_4941406 and chr05_4941464) were associated with late maturity and the haplotag C5_6_0000000 with early maturity. This QTL is near to the region containing *StCDF1* gene (Soltu.DM.05G005140.1, chr05:4485531..4488495 DMv6.1), which is well established as the gene largely responsible for the plant maturity in potato (Kloosterman et al. 2013).

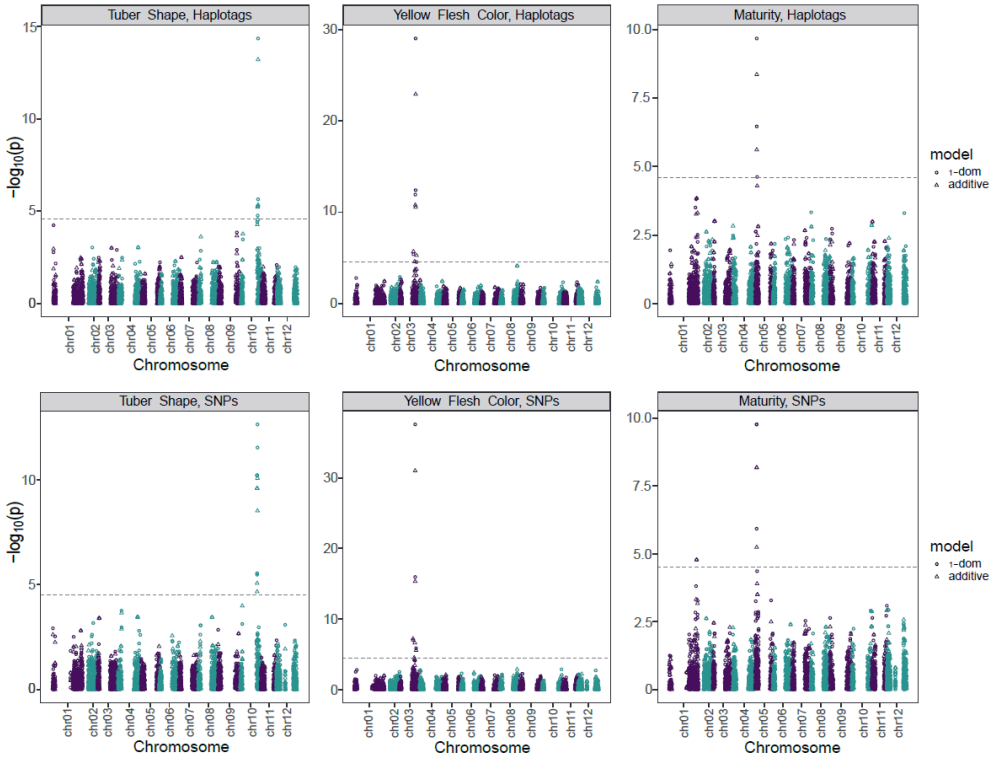


Figure 7. Manhattan plots for the reference traits: Tuber Shape, Yellow flesh colour and Maturity. Top: analysis with Haplotag data. Bottom: analysis with SNP data.

QTL for agronomic and morphological traits

18 of the QTL identified in this work were confirmed with previous QTL studies in potato at the diploid and tetraploid level (Table 4). We also detected new QTL for complex traits not yet reported before. In total, we discovered 19 novel QTL on eight chromosomes: Five QTL on chr02 – two QTL for Skin Smoothness, one for Sprout Dormancy, one for Total Tuber Number and one for Tuber Length. Four QTL on chr03 – one for Yield and one for Total Tuber Number. One QTL on chr06 for Skin Smoothness. One QTL on chr07 for Chipping Colour. Two QTL on chr08 – one for After-cooking Blackening and one for Cooking Type. Two QTL on chr09 – one for Eye depth and one for Sprout Dormancy. One QTL on chr10 for Total Tuber Number. Two QTL on chr11– one for Eye depth and one for Total Tuber Number. Three QTL on chr12 – one for Canopy stage 2 and two for Chipping Colour.

We detected a QTL for Canopy stage 2 (canopy coverage 10 weeks after planting) peaking at the same PotatoMASH locus where the well known QTL for Maturity was detected, C5_6 (4.94 Mb). The significant haplotag, C5_6_0000000 was associated with earliness and lower canopy cover while SNPs chr05_4941391 and chr05_4941406 were associated with lateness and higher canopy cover. This association between maturity and canopy type is also confirmed by the significant correlation between the phenotypic values of these two

traits (r^2= 0.41). The two additional QTL detected on chr01 and chr12 for early-stage canopy development (6 weeks after planting), couldn't be associated with plant maturity and seem to be caused by genetically independent loci affecting canopy vigour.

The novel QTL for Yield was detected in chr03, locus C3_17 (46.06 Mb). It was identified in both the full panel and in the sub-population "no Meijer" with a significant SNP chr03_46058754 and with the haplotag specific to this SNP, C3_17_011000, both associated with low yield. A QTL in this region was also detected for Total Tuber Number with the same significant SNP and haplotag both associated with low Total Tuber Number. This could be a new region associated with Yield and yield-related traits and is also supported by the significant correlation between Yield and Total Tuber Number (r^2=0.57). Two additional PotatoMASH loci on chr03, C3_7 and C3_29 were associated with low Total Tuber Number although we had not considered them separate loci in the QTL count.

Two additional novel QTL were detected for Total tuber number. One QTL on chr10 at locus C10_12 (53.13 Mb) was identified only in the full panel but not in any of the sub-populations, with a significant haplotag C10_12_00000100000 associated with low Total Tuber Number. No SNP was significant. One QTL was detected on chr11 at locus C11_12 (37.67 Mb), for the "no Meijer" sub-population only, with a significant haplotag C11_12_000100000 associated with low Total Tuber Number. No SNP was significant.

For Tuber Length, we detected one new QTL on chr02 at locus C2_4 (20.96 Mb), with the significant SNP chr02_20959691 with a small positive effect associated with shorter tubers (higher number of tubers per meter, TPM). This association is based on only 127 individuals of the "no Meijer" sub-population. This could be another new region associated with Yield and yield-related traits and is also supported by the significant correlation between Yield and Tuber Length (r^2=-0.19, high Yield correlates negatively with shorter tubers). A QTL for higher Yield was also detected in the same locus, C2_4, but in the sub-population "only Meijer" and with different SNPs/haplotags suggesting different origins of this locus.

Eye depth is a well-characterised trait and indeed we detected the well-known, large-effect QTL on chr10 in our full panel, spanning across the PotatoMASH loci C10_6 to C10_10 (4.74 to 51.2 Mb) peaking at 50.32 Mb. The deep eye (Eyd) phenotype was found to be associated with round tubers (Ro) (Li et al. 2005).The Eyd/eyd locus is located on chr10 and is closely linked with the major locus for Tuber Shape (Ro/ro). In the QTL detected here, the significant haplotags C10_8_0001111110, C10_9_011100000110110, C10_10_01010000 and SNP alleles chr10_49148246, chr10_49148293, chr10_49148305, chr10_49148316, chr10_50323151, chr10_50323153, chr10_50323225, chr10_50323244 and chr10_50323263, were all associated both with deep eyes and round tubers, being their effects consistent with the genetics known (Li et al. 2005). In the opposite direction of effect, we found C10_9_000101000100000 associated with flat eyes and long tubers. We also detected a novel QTL for Eye depth on chr11 at C11_8 (8.19 Mb), with a significant SNP chr11_8190769, associated with deep eyes. No specific haplotag was detected.

Skin Smoothness is a complex trait, and many complementary factors influence tubers' skin texture, such as soil and climate (Clark 1933). Earlier genetic studies by De Jong (1981) involved skin russeting as a phenotypic category, but in our panel no russeting phenotype was observed. In our study, only the skin texture was phenotyped, using a scoring scheme ranging from rough skin to smooth skin. Therefore, our study is the first to identify QTL for Skin Smoothness with no russeting. We detected three QTL for Skin Smoothness: Two QTL were detected on chr02: in PotatoMASH locus C2_8 (position 24.47 Mb) with the significant SNP chr02_24470953 of the haplotag C2_8_0000001000, and in locus C2_13 (position 28.96 Mb) with the significant haplotag C2_13_100010000. Both QTL were associated with smoother skin. The third QTL was found on chr06 at locus C6_3 (position 2.86 Mb), where the haplotag C6_3_0000101-00 was associated with rough skin, but no specific SNP allele underlying this haplotag was significant.

We detected a QTL for Sensitivity to Common Scab on chr06 at locus C6_18 (position 48.3 Mb). The significant haplotag C6_18_011111111110 was associated with susceptibility to Common Scab. This haplotag allele was present in only three individuals of the sub-population "no Meijer". The specific SNP for this haplotag (the only SNP not shared by the other haplotags in C6_18) was chr06_48297069, but was not statistically significant, most likely due to high missing data in this position (70%). We present this marker allele here, as a potential source for negative selection in future breeding, but further investigation needs to be done for validation.

Cooking Type is a complex trait. Previous studies revealed multiple QTL on multiple chromosomes: ch01, ch02, ch06, ch09, ch10 and ch011 (D'hoop et al. 2014; D'hoop et al. 2008; Kloosterman 2006). Our study is the first to report a QTL on chr08, at locus C8_21 (position 51.08 Mb), which was detected only for the "only Meijer" sub-population, with the significant haplotag C8_21_0000000000 associated with flouriness. No SNP was significant.

We detected five QTL for Chipping Colour measured after three storage conditions (Table 1). For Chipping Colour after storage at 8°C for 4 months before crisping, we identified one novel QTL on chr12 at locus C12_11 (position 9.26 Mb) with the full panel. The significant SNP allele, chr12_9258876, was associated with the dark colour of crisps. This SNP allele is shared by a few haplotags, and none of these haplotags was significant.

For Chipping Colour after storage at 8°C for 6 months before crisping, we detected one QTL on chr10 at locus C10_14 (position 55.08 Mb) with the sub-population "no Meijer". The significant SNP allele, chr10_55081844, with a positive effect was associated with a light, pure colour of crisps. This SNP allele is shared by two haplotags, but none of the haplotags was significant. A previous work with tetraploid clones collected from the breeding program in Teagasc (Ireland) for 'off the field' fry colour, detected a large effect QTL on chr10 peaking at 56.16 Mb in DMv6.1 (55.28 Mb in DMv4.3) (Byrne et al. 2020). Our significant SNP allele at chr10 is at 55081844 bp is approximately 1Mb distance from the one identified by Byrne et al (2020).

Three additional novel QTLs for Chipping Colour were detected for storage at lower temperature (4°C) for 6 months. One QTL mapped on chr03 at locus C3_17 (position 46.06 Mb) with the sub-population “no Meijer” with the significant haplotag C3_17_000010 associated with dark colour of crisps. Two other QTL were detected on this same locus for low Yield and Total Tuber Number, but with a different haplotag, C3_17_011000. Related to this, we found a negative correlation between Yield and Chipping Colour 2_4 (r^2=-0.23). We observed that only a small portion of the population (~4%) was heterozygous for those two alleles, possibly affecting this correlation, but we didn’t find a significant correlation between Chipping Colour and Tuber Number. The second QTL associated with a light, pure colour of crisps was detected on chr07 at locus C7_23 (position 50.94 Mb) with the sub-population “only Meijer”, specifically with the significant SNPs alleles chr07_50949343 and chr07_50949356. Those two SNP alleles are in complete LD but are dispersed in many haplotags and no haplotag resulted statistically significant. The third QTL was detected on chr12 at locus C12_9 (position 7.44 Mb) using the full panel, with the significant haplotag C12_9_0010010001100 associated with a darker colour of crisps. No SNP allele was significant.

It is useful to remember that the germplasm panel is derived from several independent commercial potato breeding programmes. We did not do an extensive analysis of each population source separately, but the fact that some QTL were exclusively discovered in one or the other of the "sub-populations", suggests that beneficial alleles in one population may have the potential to augment genetic gain in populations lacking those alleles (alternatively, in some populations, during the breeding efforts, those alleles were successfully purged, or simply never possessed some undesired effect alleles). For example: the four significant SNP alleles and haplotags, chr09_61254396, chr09_62045322, C9_21_000100010100 and C9_23_000001, all associated with the undesirable trait of deep eyes in the QTL EYE_C9_21-C9_23, were only found in the “only Meijer” sub-population. Those markers co-segregate in the same 25 individuals (~13%) and it is possible that they all originated from the same source with deep eyes. On the other hand, in the “no Meijer” population, the SNP chr09_61254396 is present in 11 (~3%) individuals, the SNP chr09_62045322 is only present in seven individuals (~2%), the haplotag C9_21_000100010100 is present in 9 individuals (2.4%) and C9_23_000001 is not present at all. This suggests that the two sub-populations don’t share the same ancestor or that the subpopulation “no Meijer” have successfully selected against this negative allele. Another example is in the significant haplotag C6_18_011111111110 associated with susceptibility to common Scab that is only present in three individuals in the “no Meijer” sub-population.

Differences in QTL detected with SNPs vs. haplotags

14 of the QTL were detected by haplotags only, 13 QTL were identified with SNP data only, and 10 QTL were discovered with both SNPs and haplotags. To gain a better understanding of the ability to detect QTL with either SNPs or haplotags, we manually re-examined all

individual QTL. We observed that in most cases of QTL detected with the haplotags only, the significant haplotag presents a specific composition of SNPs, but each individual SNP is dispersed across multiple haplotags of different SNPs compositions. In Table 5, we present four examples of this phenomenon. This is also visible when looking at the dosage effect of the markers. One example for this is the new QTL discovered for Skin Smoothness, where the significant haplotag C6_3_0000101-00 has negative effect, while none of the underlying SNPs have a significant effect nor the other haplotags composed by the same SNPs (Figure 8). In the opposite scenario of the QTL detected only with the SNP data, we observed that the significant SNP was shared in many haplotags (Table 6), which have a lower frequency in the population than the frequency of the significant SNP. To understand this phenomenon, we looked at the minor allele frequency of both SNPs and haplotags and observed that the minor allele frequency in the case of the SNPs is greater than 1% for most SNPs. When looking at haplotags, the frequency of individual haplotypes is much lower, with approx. 1200 of the haplotags have a frequency below 1% (Figure 9).

Table 5. Four examples of QTL detected by one unique haplotag and not with any of its constituent SNPs. The number of individuals carrying each SNP/haplotag allele is indicated within brackets.

Trait	Significant haplotag (number of individuals)	underlying SNPs (number of individuals)	Non-trait-associated haplotags sharing the same underlying SNPs (number of individuals)
Total tuber number	C11_12_000100000 (63)	chr11_37673981 (189)	C11_12_000101000 (5) C11_12_001100000 (11) C11_12_010100000 (1) C11_12_010101000 (138)
Skin Smoothness	C6_3_0000101-00 (9)	chr06_2681972 (549) chr06_2681990 (163) deletion_chr06_2681999 (489)	C6_3_0000100-00 (329) C6_3_000010000- (1) C6_3_0000100000 (182) C6_3_0000100001 (38) C6_3_0000110100 (29) C6_3_0011101-10 (148) C6_3_0100100000 (42)
Skin Smoothness	C2_13_100010000 (15)	chr02_28958095 (310)	C2_13_000010000 (164) C2_13_000011000 (11) C2_13_001011000 (126)
Chipping Colour 2_4	C12_9_001001000011100 (214)	chr12_7435710 (464) chr12_7435746 (531) chr12_7435801 (423) chr12_7435802 (464)	C12_9_001001001011100 (7) C12_9_001001100111110 (38) C12_9_001001100111110 (108) C12_9_001101001011100 (11) C12_9_000001000000000 (64) C12_9_000101000000000 (8) C12_9_000101010000000 (97)

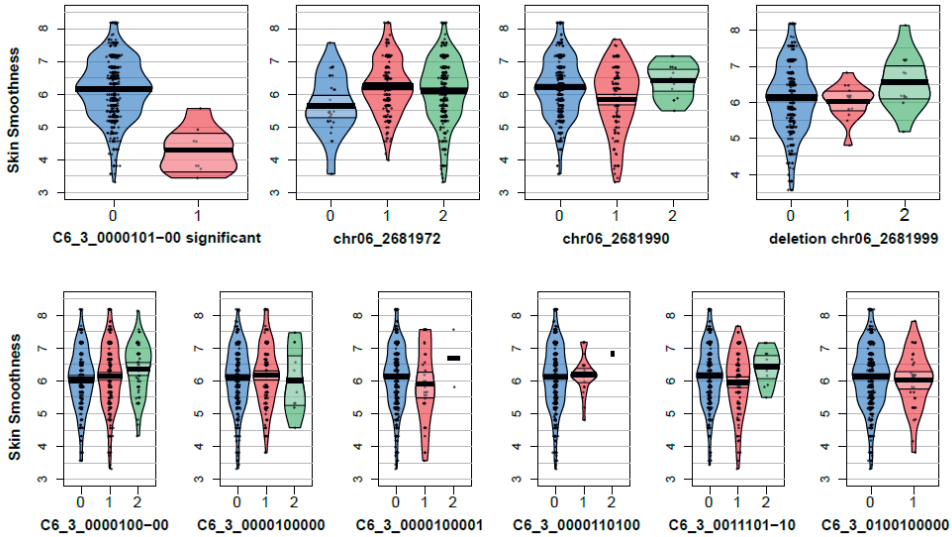


Figure 8. Allele dosage in QTL C6_3 vs the effect on Skin Smoothness. Top: Allele dosage of the significant haplotag and the non-significant SNPs underlying this haplotag. Bottom: other six haplotags of this region that possess those SNPs but their combination in the haplotags was not associated with the trait.

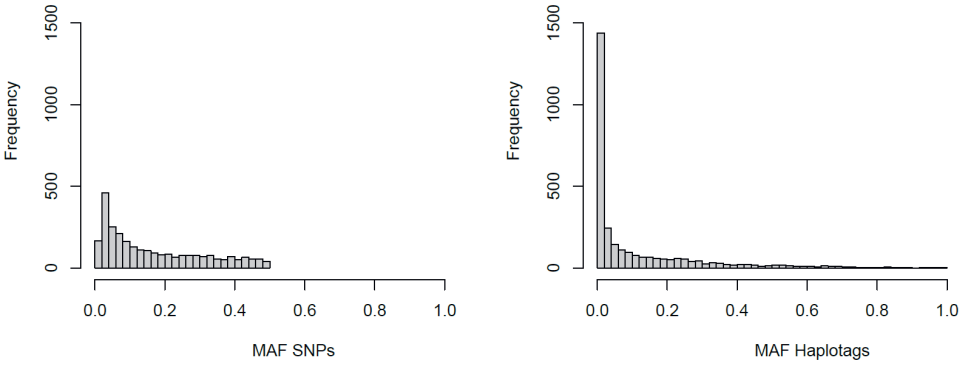


Figure 9. Minor Allele frequency (MAF) distribution of 2730 SNPs (left) and 2995 haplotags (right).

Table 6. Six QTL detected with SNPs but not with the haplotag dataset, and all the haplotags composed by those SNPs. The number of individuals for each SNP/haplotag are given within brackets.

Trait	QTL	Significant SNPs (number of individuals)	Haplotags sharing this SNPs (number of individuals)
Dry matter content	DMC C3_30	chr03_59353418 (40)	C3_30_0000100 (12) C3_30_0110100 (20) C3_30_0110110 (2) C3_30_0111100 (3)
Chipping colour 1_8	QDC1_8 C12_11	chr12_9258876 (281)	C12_13_0100000 (255) C12_13_0110000 (32)
Total tuber number	TTN C3_29	chr03_58369063 (458) chr03_58369113 (550)	C3_29_00100010 (409) C3_29_00101010 (5) C3_29_00110010 (69) C3_29_01110010 (29) C3_29_01110110 (2) C3_29_01111110 (293)
Total tuber number	TTN C3_7	chr03_37250410 (191)	C3_7_0001110 (21) C3_7_0011110 (143) C3_7_0101010 (29)
Eye depth	EYE C11_8	chr11_8190769 (239)	C11_8_00100111000 (84) C11_8_00100111100 (82) C11_8_01101111000 (73)
Chipping Colour 2_4	QDC2_4_C7_23	chr07_50949343 (155) chr07_50949356 (156)	C7_23_000111000 (71) C7_23_010010000 (2) C7_23_010110000 (30) C7_23_010110010 (45) C7_23_010111000 (3)

Discussion

Genetic improvement of potato at the diploid level is experiencing a resurgence, largely driven by the use of alleles that can overcome the gametophytic self-incompatibility system in diploid material, allowing the development of strategies to rapidly accumulate and fix traits in a manner not possible at the tetraploid level. The primary goal of this study was to genetically characterize a large pool of diploid potato breeding material that is at the foundation of the diploid breeding efforts of several commercial breeding programmes that are engaged in a collaborative initiative towards innovative potato breeding schemes, combining the analytical breeding strategy (Chase 1963), which makes use of diploids to facilitate genetic studies and selection before returning to the tetraploid level through interploidy crosses, with self-compatibility.

Phenotypic data

The diploid clones used in this study represent a very diverse collection and the commercial traits display a wide range of phenotypic trait values. In Table 2 we show that for each trait

the full scale of trait values was observed, indicative of primitive material, primary dihaploids with compromised vigour, as well as elite material. On average poor trait values were observed for quality traits such as discolouration due to Enzymatic Browning and Chipping Colour; notoriously difficult traits to improve. Most clones displayed the firm Cooking Type, which is negatively correlated with (and largely due to) low values for Dry Matter Content. Canopy development, Tuber Shape and Sprout Dormancy are among the most diverse traits. Most clones had relatively round and uniform Tuber Shape and late Maturity. Late maturity is considered beneficial to obtain an extended period of flowering, which facilitates making crosses during the breeding program, which may explain its prevalence in early stage pre-breeding material. However, early maturity is desirable for several market classes of potato. Since early maturity is largely controlled by a single large-effect quantitative trait locus, the effort to regain early maturity should be relatively easy as material advances to more commercial status over cycles of selection (Song and Endelman 2023).

Broad sense heritabilities varied mostly between moderate to high values, ranging from 50 to 90% across all traits. Sensitivity to Common Scab had, on average, the lowest H^2 across companies (50%), suggesting either low reproducibility of disease development across years due to lack of exposure to the pathogen, environmental factors, or that different trial fields used across years are infested by different isolates. This is also visible with the performance of the controls over the years and between companies, where we can see a large variance in the scoring of Sensitivity to scab even in the same year and the same company (Suppl. File 3). Moderately low average H^2 values, ranging between 60-75%, are typical for traits such as Presentability of Tubers, Skin Smoothness, Skin Brightness and Tuber Regularity, which have a somewhat ambiguous trait definition and a scale that is not objectively measurable, but rather a result of the so-called “breeders' eye". Despite the potential subjectivity of these scores, the H^2 values obtained suggest high repeatability. Genomic heritabilities were also calculated (Suppl. File 10) to catch a more accurate estimate than the average presented in Table 2. However, the results were much lower than expected, with, for example, heritability for tuber shape of 0.41 and 0.4. It could be that the tools available to calculate genomic heritability are not suited to the low number of marker sets we use in this study. This should be explored further in future analyses.

Correlation analysis of all the trait pairs was performed to examine associations between traits (Figure 2). The traits of Skin Smoothness and Skin Brightness, albeit representing a subjective breeder's score, show the highest correlation of 0.76, suggesting that the trait definitions are somewhat arbitrarily different, or share similar underlying aspects. Both traits also correlate with Yellow Skin Colour (0.48 and 0.56), where darker skins imply thicker skin. Another pair of traits: Tuber Regularity and Presentability show a Pearsons correlation of 0.66. This is not unexpected because Regularity is an aspect within Presentability, along with Eye depth (r^2 = 0.43 and 0.46). Three processing traits related to Chipping Colour also show high pairwise correlations (0.59, 0.71, 0.75), suggesting that the storage regime of tubers, causing cold sweetening, is less important than the initial Chipping quality at harvest. Byrne et al. (2020) made a similar observation on a similarly sized population of tetraploid breeding clones from a single commercial breeding programme. In this diploid gene pool,

an unexpected positive correlation of 0.57 was found between Yield and Total Tuber Number. Such a correlation would be rather unexpected for a panel of varieties, selected for yield above a certain threshold. Maturity and Canopy development also show expected correlations where late maturity leads to bigger canopies at both stages (0.34, 0.41) and the Canopy-Yield correlation was 0.25 and 0.46. However, the negative correlation between Yield and Maturity is unexpected (-0.25). Canopy stage 1 and Canopy stage 2 correlates with Total Tuber Number (0.18 and 0.37), which could be due to a common plant architecture, where stronger above- and below-ground branching patterns or stems and stolons may contribute to larger canopy cover and tuber number.

In general, this "snapshot" of the extent of phenotypic diversity in this genepool suggests that variability exists for most important agronomic and quality traits, and further selection in all or individual parts of the panel is expected to allow improvement. In terms of the use of this material in strategies involving inbreeding, we also surveyed the material for the presence of diagnostic KASP markers for the *Sli* locus as described by Clot et al, (2020), and found that *Sli* is relatively common in the material, present in 17.5% of the individuals (data not shown). The presence of this locus throughout the material means that efforts to introgress it from exotic sources, with the accompanying issues such as increased timescales of the breeding process and potential linkage drag of unfavourable loci, are unnecessary.

The germplasm panel is actually composed of material from six different breeding programmes from The Netherlands, Germany, Denmark and France. These programmes have a mixture of market class targets, including starch, table (domestic and export), processing (crisps and French fries) and speciality (e.g. salad) potatoes. Genome wide marker studies in the European cultivated potato genepool have previously shown some stratification for geographic origin of breeding programme and utility class (Uitdewilligen et al. 2015). However, this assessment is not strong, probably due to the relatively recently shared pool of progenitors of the material. When we examined the population structure of our panel, we found two highly distinct groups, one characterized by material from the breeding company Meijer and the other comprising all other material. This is interesting given the fact that diploid material from many Dutch breeding companies, including Meijer, has often originated from the diploid pre-breeding programme at WUR, whereas these groups were quite distinct in our analysis. Whilst we treated the panel as two subpopulations for some analyses on this basis, there was also some visible stratification (along the second principal coordinate in Figure 5) between the other companies. This did not correspond to either market class or geographical location of the programme.

In this experiment, different sets of diploid potato clones were grown at different locations, and trait values were evaluated by different observers. The same two or four control varieties were included in each trail to allow a fair comparison of phenotypic values. D'hoop et al. (2011) already compared phenotypic means from different experimental designs. One of their approaches made use of historical observations, on company specific candidate varieties, retrieved from breeders field books. In their other approach, all clones (now

released as variety) were grown together in a balanced trial with two locations (sandy and clay soil), both with two replications. That study showed that either a single-year balanced field trial, or multi-year-multi-location breeders' records yield robust phenotypic information that can be used in a genome-wide association study (D'hoop et al. 2011). In this study, the differences between company specific panels of diploids were controlled with structure. In particular the material offered by Meijer was also treated separately. Regarding location differences, in terms of plant material and environmental conditions, the locations may have added variance to the error term and we may have lost some power but also false negatives.

GWAS with an amplicon sequencing technique and short read haplotypes

As mentioned earlier, the primary goal of the study was to characterize the foundation breeding material that will contribute to future diploid breeding approaches focused on trait fixation through inbreeding in potato. We focused on agronomic and quality traits that breeders routinely monitor during selection in order to develop a capacity to increase the effectiveness of this process using genome-based methodologies in future. Based on the marker-inferred population structure, we preformed six QTL discovery analysis: for the entire population, the two sub-populations and with the two marker sets, SNPs and haplotags. We identified a total of 37, non-redundant QTL. Discovery of these QTL in the population gives us the potential to manipulate their configuration in future material, for instance, to accumulate and fix beneficial alleles or eliminate detrimental alleles. One potential problem with this approach in previous GWAS studies is that associated SNP markers, whilst associated with traits, may still be dispersed amongst different haplotype blocks in which the effective allele underlying the trait is also variably present. In addition to QTL discovery within this pre-breeding panel, the study also allowed us to further explore whether the multi-allelic discrimination power of PotatoMASH short read haplotypes (haplotags) can resolve this. The general approach certainly seems promising.

An average of approximately 9 haplotags was observed per PotatoMASH locus (range 2-30). This exceeds the average number we previously detected in a panel of tetraploid breeding clones (average of 6, range 2 to 14) (Leyva-Pérez et al. 2022). This greater diversity may result from the wider set of utility classes being surveyed, and because at least some of the diploid material is the result of introgression breeding for resistance loci from wild species. In our previous study on the tetraploid panel, we empirically illustrated the hypothesis that haplotags better represent the actual underlying allelic variation at a locus and may offer advantages over bi-allelic SNPs for QTL detection. We posited that this was due to better representation of regions of identity by descent harbouring the causal allelic variant of the QTL, and that haplotags are more likely to be in LD with allelic variants of genes with an effect on trait values. Conversely, some or all of the component SNPs may be dispersed across multiple haplotypes, some of which are not in LD with the effective QTL allele.

In our study, 14 of the QTL were detected by haplotags only, 13 QTL were detected with SNP data only, and 10 QTL were detected with both datasets. We found that in most cases,

these differences were due to the genetic architecture: the first situation of QTL detected with the haplotags data only, which was our original expectation, occurs when more unique haplotags are in greater LD with QTL causal alleles, whereas the bi-allelic SNPs were dispersed across multiple haplotags, some of which were not in LD with the effective QTL allele (Table 5, Figure 8).

A marker allele has to be present at sufficient frequency to support association via a statistical test. We observed much higher number of rare haplotags in our population than low-frequent SNPs (Figure 9). Therefore, for some traits, we will fail to identify statistically significant associations with haplotags with very low frequencies, which can result in the opposite phenomenon, where a QTL is only identified with SNP data. Only when a haplotag coincides with a haplotype specific SNPs their power to detect a QTL is equal.

Technical limitations affecting the power of QTL detection

The ability to detect QTL with one marker type over the other also depends on the analytical tools we use. In specific cases, we encountered that some features of our genotyping platform limited the QTL detection either with the haplotag dataset or the SNP dataset

1) PotatoMASH regions are designed to be single copy based on the DM reference genome. However, if some potato clones possess duplications with allelic variants of these regions, the reads of both copies may map back to the single copy reference genome sequence region during read mapping (since DM is also used as the reference sequence). This may be difficult to see in the SNP data, but can cause more than the expected two haplotypes in individuals with the duplication. SMAP will reject calling haplotags in these cases, assigning a missing value to that individual, and effectively filtering out instances in which this occurs. For example, in region C8_18, we detected a QTL for Enzymatic Browning in the Meijer population with the haplotag data, but not with SNP data. The SMAP locus correctness score of C8_18 was 60, suggesting the existence of additional read variation mapping to that locus (more than 2 haplotypes) that couldn't be explained as a single bi-allelic locus for 30% of the individuals. Thus, the SNP calling in this locus would be wrong, but SMAP correctly rejected calling haplotags for this locus in those individuals and the association for Enzymatic Browning in this locus is based on the remaining 60% of the population.

2) A SNP allele is significant and is in LD with one haplotag but the haplotag is not significant. Close inspection of the two QTL where this occurred showed that the significant SNP allele was present in a low number of individuals and associated as a minor effect QTL allele, for which the logarithm of the odds-score (LOD-score) was just above the threshold (Suppl. File 6). In this case, if some individuals have missing data for that locus, the haplotag cannot be called and results in a missing value. Therefore, the LOD-score of the specific haplotag may not pass the significance threshold. This could affect in two ways, either that asymmetric missing data play a role, where one allele suffers more missing data or in a symmetric way that both alleles suffer the missing data and the amount of data is simply too low to form a strong statistical test. One example for that is the SNP chr02_20959691 that was significant for Tuber Length at locus C2_4, but there was a lot of missing data for both markers sets in

this position, so the association is based on 127 individuals. The haplotag C2_4_0001000000 is specific to this SNP but was not significant; we observed that out of those 127 individuals, SMAP failed to call haplotags for 11 individuals in this locus and this probably affected the mean phenotypic score of the allelic categories, and the association that was already weak in the first place, was lost when using haplotag data.

Conclusions

Although we do not view the number of SNPs we used as “optimal” for GWAS, we were, in fact, testing the hypothesis that haplotags would detect loci not detected by the component SNPs that were used to derive them. We conclude that short read haplotags can detect additional QTL not detectable by individual SNPs, but that, for the various reasons outlined, the opposite is also true. Thus, the approach we adopted, utilising both sets of data (even though one is derived from the other) is the most optimal for QTL detection. One obstacle we faced when using haplotags for the GWAS is that we had to use them as "pseudoSNPs" to employ standard analysis software, and in this study, we have not explored the full potential of the multi-allelic nature of the haplotags. From both a genetic and practical breeding point of view, it would be interesting to gain a better understanding of the nature of allelic interactions within and between loci. Recently, Thérèse Navarro et al. (2022) developed a software, mpQTL, for QTL analysis at any ploidy level under biallelic and multiallelic models, but for multiparental populations. Their approach was demonstrated with simulated data of short-range haplotypes of autotetraploid multiparental populations. Combining approaches like this with real-world data of higher genetic diversity panel like the current study, will give insights into the genetic control of traits in highly heterozygous systems.

The increased precision offered by the new paradigm in potato breeding means that genome-based tools will become more effective in augmenting selection, allowing the "shepherding" and subsequent fixation of multiple desirable alleles into single genotypes (or the elimination of detrimental alleles). To this end, we have characterised a panel partially representative of the foundational genepool of the future of diploid breeding across several potato-breeding programmes involved in collaborative efforts in this area.

The availability of low-cost, medium-density genotyping approaches capable of generating genome-wide multi-allelic marker data in potato (e.g. PotatoMASH in this study, or the Potato DArTag EiB 1.0 generated by CGIAR https://excellenceinbreeding.org/) demonstrate that it is becoming feasible to implement such systems in breeding selection, implying the routine application to thousands of individuals per annum. These marker panels are also amenable to the addition of trait-specific markers, such as those targeting disease and pest resistance loci. We envisage a future where such assays can be used for a combination of marker assisted selection, genomic selection and monitoring the genomic constitution of inbred lines in terms of global and local homo/heterozygosity in potato breeding.

Data Availability Statement

All data necessary for confirming the conclusions of the article are present within the article's text, figures, and tables and the supplementary files.

Scripts and intermediate files for the PotatoMASH bioinformatics pipeline are also available at https://doi.org/10.6084/m9.figshare.c.6926560 .The code to reproduce the results and figures of this article are available at https://doi.org/10.6084/m9.figshare.c.6937662.

Acknowledgements

We thank the members of the public-private partnership “A new method for potato breeding: the ‘Fixation-Restitution’ approach” and SusCrop ERANET funded project “DIFFUGAT: Diploid Inbreds For Fixation, and Unreduced Gametes for Tetraploidy” (Averis Seeds B.V., Bejo Zaden B.V., Danespo A/S, Germicopa, Den Hartigh B.V., SaKa Pflanzenzucht GmbH & Co. KG, Meijer Potato, and Teagasc) for providing their support.

This research was carried out using the Teagasc high-performance computing cluster and storage systems, and the support of Dr. Paul Cormican is greatly appreciated.

Author contribution statements

D.M., H.v.E., V.P., D.G., G.L.-J, J-D.D, A.N.O.R, E.H.R.S and R.H. obtained funding, conceived and designed the study. S.W., G.L.-J, J-D.D, C.E., A.N.O.R, E.H.R.S and R.H managed the field trials, collected tissue samples and phenotypic data. L.V. isolated DNA, constructed PotatoMASH libraries, analysed genomic and phenotypic data, preformed GWAS and drafted the initial manuscript. M.d.l.O.L.-P. developed and supervised PotatoMASH molecular and bioinformatics pipeline. D.M., S.B., and A.K. advised on statistical and GWAS analysis. S.B., C.C. and T.R. advised on bioinformatics processing of the genomic data. T.R. provided the SMAP software. D.M., H.v.E., S.B., C.C. T.R. and M.d.l.O.L.-P. edited the manuscript. D.M., H.v.E. and R.G.F.V. supervised the research. All authors have read and agreed to the published version of the manuscript.

Supplementary files

Suppl. File 1: Experimental plot information.
Suppl. File 2: Correction table for the number of tubers.
Suppl. File 3: Variety controls performance.
Suppl. File 4: Normality plots for the full panel.
Suppl. File 5: Estimated means all traits for the full panel.
Suppl. File 6: Manhattan and qq plots for all traits.
Suppl. File 7: Significant markers, data with scores, effects and R^2.
Suppl. File 8: SNP dosage data.
Suppl. File 9: Haplotag dosage data.
Suppl. File 10: Marker-based heritability.

Chapter 4

Utilizing Multiplex Amplicon Sequencing and Read-Backed Haplotyping to Track Homozygosity and Residual Heterozygosity in Diploid Potato Breeding

Authors

Lea Vexler[1,2,3], Dan Milbourne[1]*, Ronald C. B. Hutten[2], Vanessa Prigge[4],Christel Engelen[2], Stephen Byrne[1], Maria de la O Leyva-Perez[1], Denis Griffin[1], Richard G.F. Visser[2], Herman J. van Eck[2]

Affiliations

[1]Teagasc, Crop Science Department, Oak Park, R93 XE12 Carlow, Ireland

[2]Plant Breeding, Wageningen University & Research, P.O. Box 386, 6700 AJ Wageningen, The Netherlands

[3]Graduate School Experimental Plant Sciences, Wageningen University & Research, Wageningen, The Netherlands

[4]SaKa Pflanzenzucht GmbH & Co. KG, Eichenallee 9, 24340 Windeby, Germany

Abstract

Diploid potato breeding is frequently challenged by residual heterozygosity (RH), which impedes the generation of pure inbred lines for F1 hybrid production. Fixation-Restitution Breeding strategy (Fix-Res), which employs RH to optimize beneficial traits, is a promising alternative. Maintaining heterozygosity throughout the inbreeding process is crucial to its success. In this study, we used a multiplex amplicon-sequencing assay (PotatoMASH), in conjunction with read-backed haplotyping, as an effective tool for tracking homozygosity in diploid breeding. Utilising a collection of 271 inbred diploid clones from the Wageningen diploid breeding program, we obtain a "snapshot" of the genetic composition shaped by over 40 years of breeding efforts. Furthermore, we examine a self-compatible individual lineage from the program to identify key hotspots of homozygosity across chromosomes. Our results demonstrate that multiplex amplicon sequencing with read-backed haplotyping, through haplotag construction, provides a more accurate and reliable measure of homozygosity compared to traditional SNP-based methods, without the requirement for parental genomic data. In S3 progenies, we observed average homozygosity levels of 82-83.4% for Identity-by-State (IBS) and 72.6-74.8% for Identity-by-Descent (IBD), which were lower than anticipated. Hotspots of heterozygosity were detected across all chromosomes, with chromosome 5 entirely heterozygous across three generation of selfing and chromosome 11 reaching full homozygosity in one S3 progeny. This absence of homozygosity on chromosome 5 may be associated with reproductive-related QTLs and genes that favour the heterozygous state due to potentially deleterious alleles. These findings demonstrate the potential of assays such as PotatoMASH to effectively track and optimize heterozygosity in Fix-Res breeding programs.

Introduction

The conventional breeding scheme for developing improved potato varieties primarily targets tetraploid cultivars, relying on clonal selection of segregating offspring from two highly heterozygous parents. However, beneficial allele combinations from the parental genotypes often disassemble during meiosis, and recombination of favourable alleles in the offspring becomes a challenging probabilistic process, leading to slow genetic gains over time. In contrast, diploid breeding can potentially simplify this process due to less complex allelic combinations, enabling more efficient selection against undesirable alleles. Thus, diploid breeding offers a promising alternative to conventional tetraploid breeding by enhancing both the efficiency and speed of genetic progress.

In many crops, the development of inbred lines for diploid breeding is routine, but in diploid potato, breeders face two major challenges. The first obstacle is the self-incompatibility of most diploid clones, a problem that has been the focus of extensive research (De Jong and Rowe 1971b; Hosaka and Hanneman 1998b; Hosaka and Hanneman 1998a; Lindhout et al. 2011; Olsder and Hermsen 1976; Phumichai and Hosaka 2006). The identification of the *Sli* locus at the distal end of chromosome 12 by Hosaka and Hanneman (1998a, 1998b) was a breakthrough in understanding self-compatibility in potatoes. More recently, Clot et al. (2020) found that the *Sli* haplotype is widespread in both tetraploid varieties and diploid cultivated germplasm and developed KASP markers specific to this haplotype, which have been rapidly adopted by the potato genetics community (Kaiser et al. 2021; Song and Endelman 2023; Sood et al. 2024). The second and more challenging hurdle is inbreeding depression, which occurs when deleterious recessive alleles are exposed during self-fertilization, resulting in a loss of vigour and fertility. This limits the potential for continuous selfing in potatoes (Charlesworth and Willis 2009; De Jong and Rowe 1971b; Krantz 1924; Krantz 1946).

Over a decade ago, the F1 hybrid breeding scheme was introduced for potato (Lindhout et al. 2011), and since then several research groups worldwide have initiated programs to enhance the effectiveness of recurrent selection by breeding potato at the diploid level (Bradshaw 2022; Song and Endelman 2023; Sood et al. 2024; Stokstad 2019; Zhang et al. 2022b; Hosaka and Sanetomo 2020; Jansky et al. 2016), utilizing self-compatible diploids in F1 hybrid breeding. For this breeding method, highly inbred parental genotypes are hybridized to produce uniform F1 botanical seed as propagating material, with any residual inbreeding depression being dealt with by exploiting the effect of heterosis (De Jong and Rowe 1971a; Lindhout et al. 2011). However, initiating F1 hybrid breeding in potatoes requires the development of near or complete homozygous inbred lines. While higher homozygosity can be achieved through several generations of selfing, studies show that advanced selfed generations suffer from inbreeding depression, resulting in reduced vigour, slower growth, and fertility-related issues, such as poor flowering (Hosaka and Sanetomo 2020; Peterson et al. 2016; Phumichai and Hosaka 2006; Zhang et al. 2019; Wu et al. 2023).

On the contrary, maintaining some level of heterozygosity has been shown to improve traits such as yield, tuber number, and reproductive performance, with fertility selection favouring more heterozygous plants (Marand et al. 2019; Phumichai et al. 2005; Peterson et al. 2016). This suggests that a certain level of heterozygosity might be essential for maintaining fertility and minimizing inbreeding depression. However, this concept clashes with the principles of F1 hybrid breeding, where highly inbred parental components are required to produce uniform F1 true seeds. This divergence highlights the ongoing challenge of balancing the need for inbreeding depression control with the requirement for homozygosity in potato TPS (True Potato Seed) parent breeding programs.

Clot (2023) presents a novel strategy called Fixation-Restitution Breeding (Fix-Res Breeding). This approach captures many of the benefits of diploid potato breeding but makes the process compatible with breeding at the tetraploid level. In the first step, similar to diploid F1 hybrid breeding, the diploid potato is rendered self-compatible by introgression of the single dominant self-compatibility gene (*Sli*), allowing inbreeding and recurrent selection to accumulate and fix beneficial alleles. Unlike traditional F1 hybrid breeding, the diploid potato's ability to produce unreduced gametes through chromosome restitution is utilized in the second step, where interploidy crossing between the diploid and a tetraploid occurs, ensuring a complete transfer of the chromosomal content from the diploid to the tetraploid progeny population. This feature is a significant advantage of Fix-Res Breeding, as it ensures the preservation of beneficial alleles from the diploid parent, even if they are in a heterozygous state. By maintaining higher heterozygosity, this strategy promotes traits that benefit from genetic diversity. Ultimately, it combines the advantages of heterozygosity with the compatibility of tetraploid breeding, making it a promising approach for potato improvement.

In diploid potato breeding, accurately tracking heterozygosity throughout the process is useful regardless of the chosen breeding strategy. Despite potatoes being known for their high heterozygosity, limited published data quantifies this trait at both the diploid and tetraploid levels. With the rise of next-generation sequencing technologies and available SNP data, an increasing number of studies have begun genotyping larger panels of potato clones to estimate heterozygosity. Methods for estimation vary, with some using SNPs as direct markers (Lindhout et al. 2011; Peterson et al. 2016; Leisner et al. 2018; Hosaka and Sanetomo 2020; Jansky et al. 2014), while others align SNPs into bins or generate parental haplotypes (Song and Endelman 2023; Zhang et al. 2021; van Lieshout et al. 2020; Marand et al. 2019). Some studies report high levels of homozygosity achieved through generations of selfing, such as 97% homozygosity in S5 plants (Zhang et al. 2021), 90% and 95% homozygosity in diploid S7 clone M6 (Leisner et al. 2018; Jansky et al. 2014), and even complete homozygosity in S10 and S11 plants (Hosaka and Sanetomo 2020). However, these findings contrast with reports of residual heterozygosity (RH). For instance, a recent study on the self-compatible diploid S9 clone “Solyntus” found that 20% of its genome remained heterozygous (van Lieshout et al. 2020).

These discrepancies may stem from differences in how heterozygosity is measured and the lack of standardized methods for estimating it in potato populations. A key distinction is whether homozygosity is assessed based on identity-by-state (IBS) or identity-by-descent (IBD). IBS refers to alleles that are identical in sequence regardless of their ancestral origin, while IBD refers to alleles that are both identical and inherited from a common ancestor (Powell et al. 2010). In the context of homozygosity estimation, IBS assumes that loci appearing homozygous in the parent are inherited identically in the offspring, which can lead to overestimation. In contrast, IBD-based approaches may exclude such loci unless both alleles are confirmed to be inherited from the same ancestral source, potentially underestimating homozygosity. The choice between these two approaches, and the presence or absence of parental information, may therefore underlie the observed inconsistencies in heterozygosity estimates across studies. This highlights the need for standardized methods that consider data availability and study design to ensure comparability and accuracy in breeding research. In the present study, we applied both IBS and IBD homozygosity estimates during our analysis, which allowed us to determine which approach was more reliable given our dataset and breeding design.

In a previous study (Leyva-Pérez et al. 2022), we developed PotatoMASH, a high-throughput, cost-effective marker system designed for genome-wide genotyping in potato populations. PotatoMASH utilizes a multiplex amplicon sequencing approach followed by deep next-generation sequencing (2x150bp Illumina), targeting 339 loci and generating over 2000 SNPs due to the high SNP density in potato germplasm. Additionally, tools for read-backed phasing (Schaumont et al. 2022) are used to generate short haplotypes (165-180bp) that capture allelic diversity at each locus. Designed to be cost-effective, PotatoMASH is suitable for routine use in potato breeding applications. In a previous study, we demonstrated the system's efficiency in diagnosing targeted pest-resistance markers (Leyva-Pérez et al. 2022), detecting multiple quantitative trait loci (QTLs) through genome-wide association studies (GWAS), and tracking variation in a diploid segregating population (Leyva-Pérez et al. 2022; Vexler et al. 2024). In the present study, we aim to evaluate the effectiveness of PotatoMASH, specifically the short-read haplotypes (haplotags), in assessing changes in genome-wide homozygosity and inferring selfing rates in breeding material. A key focus of this research is to trace RH regions, which are critical for the preservation of beneficial alleles that may be maintained in repulsion with deleterious alleles (Jansky et al. 2014). This analysis will play a crucial role in the development of Fix-Res recurrent parents, ensuring that favourable allelic combinations are retained during the breeding process, thereby improving the efficiency and outcomes of potato breeding programs.

Almost forty years ago, the Institute for Plant Breeding (IVP) at Wageningen University initiated its diploid potato breeding program. The genetic composition of this material was derived from *S. tuberosum* (about 50%), *S. phureja* (about 40%), and more than 10 wild species (about 5-10% in total). In the current study, we aimed to analyse the genetic composition of clones derived from this long-term diploid breeding program. Additionally, we

focused on a self-compatible lineage from the program that underwent recurrent self-pollination up to the S3 level through single seed descent. The objective was to assess the extent and distribution of homozygosity across the selfing generations and identify regions of persistent RH, which are likely critical for maintaining the viability, fertility, and vigour of selected traits throughout the breeding process.

Materials and methods

Plant materials

The diploid breeding germplasm

A diverse sample of 271 clones was selected from the diploid potato breeding program at Plant Breeding, Wageningen University & Research (PBR-WUR), representing different breeding stages: 198 parents, 60 S1 individuals, 10 S2 individuals, and 3 S3 individuals (Suppl. File 1). The inbred program at PBR-WUR follows a 2-3 year generation cycle:

Year 1: Selfing of clones

Year 2: Raising of true seedlings and visual selection of vigour and tubers

Year 3: The best true seedlings from Year 2 are planted in the crossing greenhouse for further selfing.

Each clone progresses through this cycle individually. While selfing occurs in both Year 1 and Year 3, these refer to different generations within the same lineage — i.e., Year 3 marks the beginning of the next selfing generation.

Implicit to this approach is that parental clones are self-compatible and have good male and female fertility. In addition to selfing, various crosses are made, such as backcrosses (P × S1), hybridisations of related/unrelated first generation (S1 × S1), or second generation inbreds (S1 × S2). While most inbred clones performed well for certain traits inherited from their parents, none exhibited strong performance across all traits. In cases where vigorous and well-tuberizing inbreds were obtained but struggled to self, these clones were backcrossed to the parent or with other selfing-capable clones. This strategy was employed because we believe such crosses offer a better starting point than strict selfing to proceed with this program. The aim of the program is the development of fertile and self-compatible, partially inbred clones that have already reached fixation of more beneficial alleles as present in the parent.

For the purpose of this study, 198 individuals are considered parents, even though they are somewhat inbred and are actually crosses of related parents, meaning their expected homozygosity will be higher than zero. The diploid program is represented as a pedigree network in Suppl. File 1.

Experimental progenies

Next to the unstructured panel of inbreds from the breeding program we developed experimental offspring from clone IVP92-030-14, to generate 173 inbred offspring, including 37 S1 individuals, 48 S2 individuals, and two S3 families, with 41 and 48 individuals each, as outlined in Figure 1.

For clarity, the 271 individuals from the PBR-WUR breeding program will be referred to as the diploid breeding germplasm throughout the study, while the inbred offspring of IVP92-030-14 will be referred to as the experimental progenies.

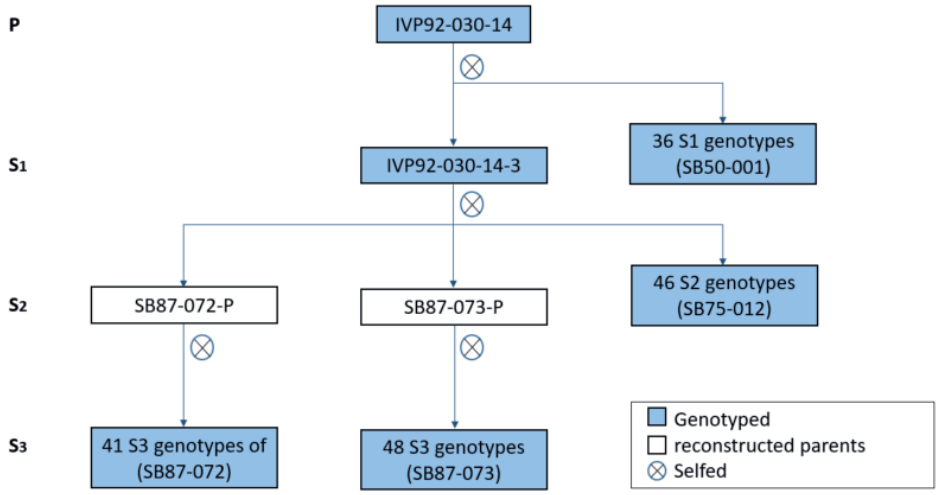


Figure 1. Pedigree of the experimental progenies derived from IVP92-430-14. The parental individuals of the two S3 families were not maintained. Based on the genotypes of their offspring the genetic composition of SB87-072-P and SB87-073-P was reconstructed.

Genotyping with PotatoMASH

Leaf material was sampled, freeze-dried and stored with silica gel until use. Approximately 5 mg of dry tissue was used to extract DNA with Mag-BIND® Plant DNA DS Kit (Omega-VWR M1130-00, Philadelphia, USA), using the KingFisher Flex automated extraction & purification system (Thermo Scientific, Austin, TX, USA).

PotatoMASH libraries construction and SNP calling was performed as in Leyva-Pérez et al. (2022) (https://www.protocols.io/view/potatomash-library-construction-e6nvw53zdvmk/v2) with the final adjustments to the bioinformatics pipeline as in Vexler et al. (2024) using the DMv6.1 reference genome (Pham et al. 2020).

For haplotag calling, we utilized the SNP-based haplotag calling approach with the haplotype-sites function of the SMAP software (Schaumont et al. 2022), combined with discrete genotype calling, which were then used as presence-absence markers.

Data analysis

All statistical analyses, data mining and visualizations were performed using R version 4.2.1 unless otherwise specified in results.

Admixture with single foreign pollen grain can distort our analysis, and therefore a paternity test was performed on the experimental progenies using the haplotag data before proceeding with further analyses. Three individuals were identified as erroneous due to marker scores, indicating that they cannot be offspring from the assumed parent and were removed from further analysis.

For the S3 individuals in the experimental progenies, the parental clones SB87-072-P and SB87-073-P were not available for genotyping, so the genotypic data for these parents were inferred based on the offspring's genetic information. For loci that showed no segregation (i.e., present in at least 95% of the offspring), we assumed the parents' genotype to be homozygous and identical to that of the offspring. For loci exhibiting segregation, we inferred the parents' genotype to be heterozygous. Markers with more than 50% missing data were considered missing for the corresponding parent.

Analyses were conducted in terms of observed and expected homozygosity. The observed homozygosity was calculated by differentiating between Identity By State (IBS) homozygosity and Identity By Descent (IBD) homozygosity, as specified below.

For each individual clone, the number of homozygous (dosage 2 for haplotags, and both dosage 2 and 0 for SNPs) and heterozygous (dosage 1) loci were counted. The observed IBS homozygosity was estimated as the proportion of homozygous loci to the total number of genotyped loci. The observed IBD homozygosity was estimated using a subset of loci that were heterozygous in the parent. The IBS estimates represent the upper boundary and the IBD estimates the lower boundary of the true homozygosity.

Expected homozygosity for inbreds was derived from a simple Mendelian model, assuming no selection. In this model, the number of heterozygous loci was expected to decrease by 50% in each generation. A chi-square test ($\alpha = 0.05$) was used to test the null hypothesis of no difference between the observed and expected values for each individual.

In the experimental progenies, the mean and standard deviation of the observed homozygosity were calculated for each inbred family. A t-test was then used to assess the null hypothesis of no difference between the family mean observed homozygosity and the expected homozygosity, in order to evaluate deviations from the expected level of inbreeding in each family.

Population structure was evaluated using principal component analysis (PCA), performed with Plink 1.9 using SNPs with a Minimum Allele Frequency >0.01 (Purcell et al. 2007). A scatterplot was generated with 95% confidence ellipses, assuming multivariate t-distribution.

The pedigree network (Suppl. File 1) was constructed using Helium v. 2.0.0 software (Shaw et al. 2014).

Results and Discussion

Monitoring homozygosity levels and identifying heterozygous regions throughout the development of inbred lines is essential in diploid breeding programs. This process is crucial for both creating highly homozygous inbred lines for F1 hybrid potato breeding as well as producing partially inbred diploid clones as recurrent parents in Fix-Res breeding. In this study, we genotyped two distinct groups of diploid clones. The first group included 271 diploid clones from the PBR-WUR diploid breeding program, referred to as "The diploid breeding germplasm". This group offers a comprehensive snapshot of the genetic composition resulting from over 25 years of continuous diploid breeding efforts. The second group consisted of 174 individuals derived from a single lineage of a self-compatible individual, hereafter referred to as "The experimental progenies". This allowed us to investigate variations in homozygosity across three generations of selfing and identify regions of persistent heterozygosity throughout the inbreeding process of this particular lineage, which was frequently used in the diploid breeding program due to its high fertility and stable self-compatibility, as its parent was homozygous for the *Sli* gene.

Genotyping, variant calling, and genetic diversity

We conducted genotyping on a total of 442 individuals from both the diploid breeding germplasm and the experimental progenies using PotatoMASH multiplex-amplicon sequencing (Leyva-Pérez et al. 2022). Following the merging and filtering of the sequencing reads, we retained an average of 889,734 reads per potato clone, corresponding to approximately 1,450 reads per amplicon per sample. This read depth significantly exceeds the recommended minimum of 20x for SMAP haplotype calling in diploid species. The amplification efficiency of the primer pairs, whether low or high, remained consistent across most samples (Figure 2).

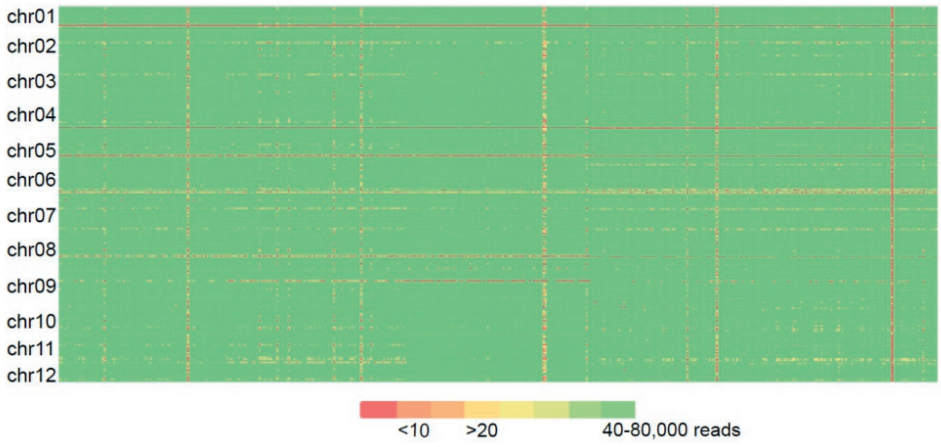


Figure 2. Read coverage of the 339 PotatoMASH loci. Heatmap showing the number of merged and filtered reads from 442 samples (represented by columns) that mapped to each of the loci (represented by rows).

Following filtering, 2,315 SNPs were identified across the panel. Of the 339 target loci, SMAP successfully identified 333 multi-allelic loci. A total of 1,823 haplotags were detected across the panel, with each locus containing between 2 and 18 haplotags, and a mean of 5.5 haplotags per locus (Table 1, Figure 3).

Table 1. Summary of genotyping and variant calling with PotatoMASH.

Total samples	**442**
SNPs called	4,971
SNPs filtered	2,315
Polymorphic loci	333
Number of haplotags	1,823
Haplotypes per locus	2-18 (5.5

The Minor Allele Frequency (MAF) profile of the two datasets, SNPs and SNP-based haplotags, exhibited a similar distribution, although the haplotag dataset contained a slightly higher proportion of loci with rare alleles (low MAF) (Figure 3).

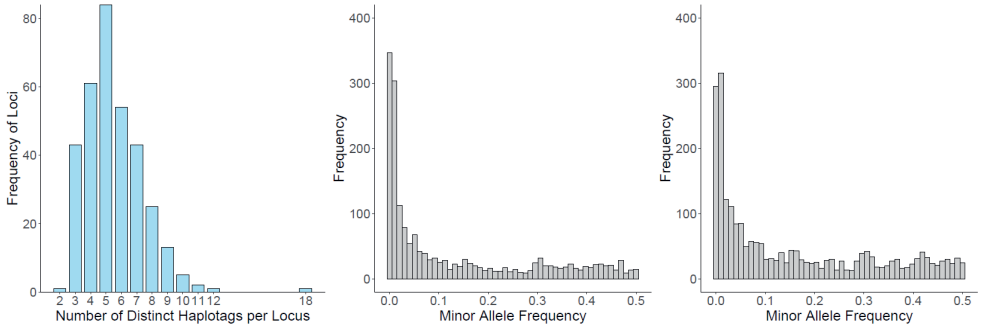


Figure 3. (a) Distribution of haplotag diversity across 333 loci in all individuals in the study. (b) Minor Allele Frequency (MAF) distribution for 2,315 SNPs. (c) MAF distribution for 1,823 SNPs-based haplotags.

We analysed the population structure using Principal Component Analysis (PCA) based on SNP data (Figure 4). Initially, we assessed the overall population structure of all 442 individuals in this study (Figure 4a). The first two principal components explained a substantial proportion of the genetic variance, with PC1 accounting for 35% and PC2 for 8%. The experimental progenies formed a distinct cluster within a smaller region of the spectrum. When we analysed the population structure separately for the diploid breeding germplasm and the experimental progenies (Figures 4b and 4c, respectively), we found that the diploid breeding germplasm exhibited significantly greater genetic diversity, with PC1 and PC2 explaining 19% and 11.12% of the genetic variation. In contrast, the experimental progenies displayed a more pronounced genetic structure, with each selfing generation forming distinct clusters. The genetic pool in this group was more restricted, resulting in the

first two principal components (PC1 and PC2) explaining a larger proportion of the overall variance (36.33% and 17.3% respectively).

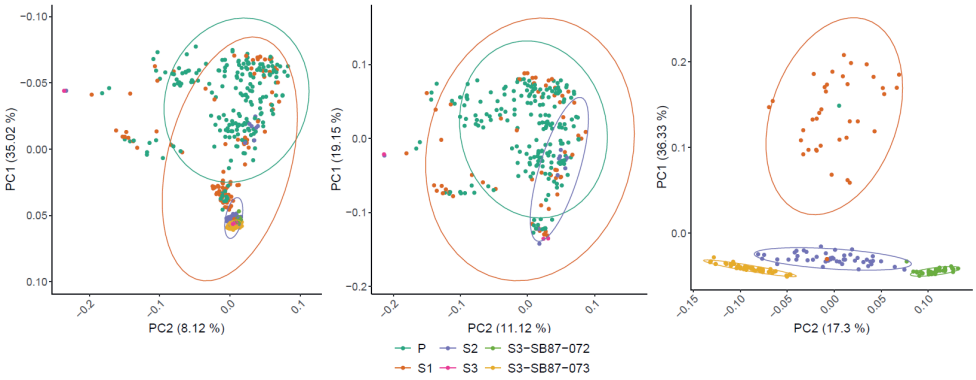


Figure 4. Population structure with Principle Component Analysis (PCA) of the SNP data with 95% confidence ellipses. (a) PCA for the full collection used in this study. (b) PCA of the diploid breeding germplasm. (c) PCA for the experimental progenies. Labels: P = parental clones, S1 = first selfed generation, S2 = second selfed generation, S3 = third selfed generation. S3-SB87-072 and S3-SB87-073 refer to two distinct S3-derived clones within the experimental population.

Tracking homozygosity in germplasm from a long running diploid breeding programme

A diverse sample of 271 clones from the diploid breeding program was assessed for homozygosity levels, comprising 198 founder clones (referred to as "parents"), along with 60 S1, 10 S2, and 3 S3 individuals, constituting three generations of selfing from different founders (Suppl. File 1).

Comparison of homozygosity measurements based on SNPs and haplotags

Homozygosity was measured across generations of selfing in the diploid breeding germplasm using two distinct marker datasets: SNPs and SNP-based Haplotags. Detailed observed homozygosity values for each individual are presented in Suppl. File 3, with a summary of the results in Table 2. Due to the absence of parental genomic data for most clones, which were not maintained long-term due to storage and propagation constraints, we initially calculated Identity By State (IBS) homozygosity by counting all homozygous markers. For the SNP-based analysis, SNP loci in a homozygous state for either the reference allele (dosage 0) or the alternative allele (dosage 2) were considered homozygous. With a total of 2,315 SNPs, many of which exhibited low frequencies in the population (Figure 3b), many SNPs were homozygous for the reference allele, leading to inflated homozygosity estimates. This resulted in high average homozygosity levels: the parents showed an average of 75.7%, which increased to 83.1% in S1, 87.8% in S2, and 90.4% in S3 (Table 2).

In contrast, the haplotag-based approach provided more accurate estimates of homozygosity. Homozygosity was calculated by counting loci homozygous for one haplotag

(dosage 2). Utilising this approach, the average observed homozygosity for the parents was 41.8%, which increased to 59.1% in S1, 71.8% in S2, and 77.2% in S3 (Table 2).

Table 2. Summary of Identity by State (IBS) homozygosity results for the 271 individuals in the diploid breeding program across two marker datasets (SNPs and Haplotags). The table includes the number of individuals, expected homozygosity values, average observed homozygosity, range (min-max), and standard deviation (St.Dev) for each group (Parents and successive selfing generation: S1, S2, S3).

	Parents (Founders)	**S1**	**S2**	**S3**
Number of individuals	198	60	10	3
Expected homozygosity*	0%	50%	75%	87.5%
SNPs				
Average Observed homozygosity	75.7%	83.1%	87.8%	90.4%
Range Observed homozygosity (Min-Max)	69.1-86.3%	76.7-89.6%	82.2-91.3%	87-92.9%
St.Dev	5.7%	4.3%	3.2%	2.2%
Haplotags				
Average Observed homozygosity	41.8%	59.1%	71.8%	77.2%
Range Observed homozygosity (Min-Max)	21.8-68.2%	42.1-75.3%	64.3-78.4%	74.8-78.8%
St.Dev	15.1%	11.2%	5.5%	1.5%

**Expected homozygosity was derived from a simple Mendelian model, assuming no selection.*

These results highlight the advantages of the haplotag-based approach, which enhances allele resolution and mitigates false positives caused by "identity-by-state" without necessitating parental genomic data.

These findings are in line with the very high homozygosity levels previously reported, as outlined in the introduction. For example, Hosaka et al. (2020) used a set of 18.5k SNPs and reported complete homozygosity in S10 and S11 plants. While the authors acknowledged the possibility that some heterozygous regions might remain undetected, they also reported 80% homozygosity in F1 plants, which is notably high. This suggests that their observations may have been influenced by false positive homozygosity due to IBS in SNP data. We propose that haplotag data provides a more reliable estimate of homozygosity

compared to SNP data, and it serves as an effective tool for screening homozygosity in partially inbred clones, without the requirement for parental genomic information.

Analysis of IBS and IBD homozygosity in reference to parental genotypes

When parental genomic information is available, different approaches can be taken to assess homozygosity in breeding programs. In the present analysis, 34 of the 271 parents served as progenitors of 67 partially inbred clones, including 56 S1 plants, 10 S2 plants, and 1 S3 plant. The availability of parental genomic information allowed us to calculate homozygosity using both IBS and IBD approaches for SNP and haplotag datasets. As outlined in the introduction, IBS and IBD differ in how they interpret shared alleles, which can lead to overestimation or underestimation. Given the nature of these two approaches, the true homozygosity value likely falls between the IBS and IBD-based estimates. Detailed homozygosity values for each individual are provided in Supplementary File S4, with a summary of results in Table 3.

With the availability of parental genotype data, we were also able to exclude markers that were non-polymorphic in the parents and calculate both IBS- and IBD-based homozygosity for both SNP and haplotag datasets. After this filtering step, we observed a reduction in the estimated SNP-based IBS homozygosity compared to earlier findings. For the SNPs, the average observed homozygosity for the parents was 41.4%, increasing to 59.2% in S1, 71.7% in S2, and 70.3% in S3. These values were comparable to those derived from haplotag data, where the average observed homozygosity for the parents was 43.1%, increasing to 59.8% in S1, 72.0% in S2, and 75.5% in S3.

For IBD homozygosity, which focuses on heterozygous loci in the parents, lower values were observed compared to IBS. Using SNP data, the average IBD homozygosity in S1 was 34.8%, increasing to 55.7% in S2. The only individual in S3 exhibited 58.6% IBD homozygosity. Similarly, the haplotag data showed trends with an average IBD homozygosity of 33.9% in S1, increasing to 55.8% in S2, and 64.4% in S3.

We also calculated the expected homozygosity for each generation of selfing based on Mendelian inheritance, with a 50% reduction in heterozygosity for each generation. Chi-square tests on genotypic counts revealed significant deviations from these Mendelian expectations for most offspring, suggesting that observed homozygosity was often significantly lower than expected (Suppl. File 4). Although we present these results in Table 3, the unstructured panel of inbred subsets (with small sample sizes in S2 and only one S3 plant) made it challenging to discern clear trends between observed and expected homozygosity. Therefore, further analysis was conducted using the experimental progenies, which were more suitable for this type of analysis.

Table 3. Summary of Identity by State (IBS) and Summary of Identity by Descent (IBD) homozygosity results for the 67 individuals in the diploid breeding program where the parents were also genotyped, across two marker datasets (SNPs and Haplotags). The table includes the number of individuals, expected homozygosity values, average observed homozygosity, range (min-max), and standard deviation (St.Dev) for each group (Parents and successive selfing generation: S1, S2, S3).

	IBS				IBD			
	Parents	**S1**	**S2**	**S3**	**Parents**	**S1**	**S2**	**S3**
Number of individuals	34	56	10	1	34	56	10	1
SNPs								
Expected homozygosity	-	68.9%	78%	79.1 %*	0%	50%	65.5%	71% *
Average Observed homozygosity	41.4%	59.2%	71.7%	70.3 %*	0%	35%	55.70 %	59.1 %*
Range Observed homozygosity (Min-Max)	29.4-60.8%	44.3-71.3%	58.2-78.6%	-	0%	18.7-53.2%	42-67.7%	-
St.Dev	8.7%	8.9%	7.2%	-	-	11.5%	9.2%	-
Haplotags								
Average Expected homozygosity	-	69.9%	77.5%	82.2 %*	0%	50%	64.6%	74.1 %
Average Observed homozygosity	43.1%	59.8%	72%	75.5 %*	0%	33.9%	55.8%	64.4 %*
Range Observed homozygosity (Min-Max)	21.8-64.3%	41.9-75.3%	64.5-78.5%	-	0%	16.9-55.7%	43.8-66.8%	-
St.Dev	10.1%	11.3%	5.4%	-	-	12.7%	8.6%	-

**Based on one individual*

With the availability of parental genomic data, our analysis demonstrated no clear advantage to using SNPs over haplotags, as both marker types yielded comparable results for estimating homozygosity with either the IBS or IBD approach. Consequently, we focused on haplotag data for further analysis, with SNP-based results provided in the supplementary material. When using haplotags, we observed substantial differences between homozygosity estimates from IBS and IBD, with true homozygosity likely falling between the two. The choice between IBS and IBD approaches should be guided by the available data and study objectives. While IBD provides insights into selfing rates and homozygosity evolution over generations, it may underestimate true homozygosity levels, as it does not account for the initial homozygosity level of the parents. Additionally, IBD calculations require parental genotypes, which are not always available. In such cases, IBS can be used as an alternative, though it may overestimate homozygosity due to its inability to distinguish between IBD and IBS alleles.

Using haplotags to track homozygosity levels in an experimental inbred population

In the previously described analysis, we examined the allelic state of individual haplotags. Building upon this investigation, we sought to explore variations in homozygosity across generations of selfing and identify chromosomal regions where heterozygosity persists through inbreeding. To achieve this, we designed an experiment using a self-pollinating founder clone homozygous for the *Sli* locus, IVP92-430-14, from the inbred diploid program at PBR-WUR. This founder was utilized to generate experimental progenies to study the increase in homozygosity levels through successive generations of selfing. As outlined in the methods (Figure 1), the founder was selfed to produce an S1 progeny consisting of 37 offspring. One individual from the S1 progeny (IVP92-030-14-3) was then selfed to produce an S2 progeny of 48 offspring. Two individuals from the S2 progeny were selfed to generate two S3 progenies: S3-SB87-072 and S3-SB87-073. Since the two S2 individuals were not maintained, their genotypic data were reconstructed from their offspring. Here, we present the results obtained using haplotag data, with results from SNP data provided in Suppl. File 5.

The dataset includes 327 PotatoMASH loci for the founder, with 215 heterozygous and 112 homozygous loci. All loci were used to calculate IBS homozygosity, resulting in a starting homozygosity of 34.3% for the F1 founder IVP92-030-14. For the IBD homozygosity calculation, only heterozygous loci were considered, and the founder's starting homozygosity was set to 0%.

Table 4 summarizes the results. A Chi-square test was applied to the genotypic counts to evaluate deviations from Mendelian expectations for each selfed offspring (Suppl. File 5). A broad range of homozygosity estimates was observed across individuals within all progeny families.

Homozygosity increased from parent to progeny across all progeny types (Table 4, Suppl. File 6). When calculating homozygosity using the IBS approach, the average increase from the founder parent to the S1 progeny was 27.1%. The S2 progeny showed an average increase of 17.3% from the S2 parent, and the S3 progenies (S3-SB87-072 and S3-SB87-073) showed average increases of 10.6% and 14.6%, respectively. When calculating homozygosity using the IBD approach, the average increase from the founder parent to the S1 progeny was 41.3%. For the S2 progeny, the increase averaged 26.5%, and the S3 progenies demonstrated average increases of 15.7% (S3-SB87-072) and 21.9% (S3-SB87-073).

Overall, statistically significant differences were observed between the observed and expected mean progeny homozygosity levels, with all t-tests demonstrating significance differences. The average expected homozygosity was consistently lower than predicted by the Mendelian model. Smaller discrepancies between observed and expected homozygosity were observed as progeny inbreeding increased. With IBS, the differences were 6% for S1,

4% for S2, and 3%-2% for S3. With IBD, the differences were 9% for S1, 7% for S2, and 5%-3% for S3 (Table 4).

Table 4. Summary of homozygosity results for 174 selfed progeny individuals using the haplotags, analysed for Identity by State (IBS) and Identity by Descent (IBD). The table includes the number of individuals per progeny type (S1 = first selfed generation, S2 = second, S3 = third. S3-SB87-072 and S3-SB87-073 refer to two distinct S3-derived clones), the parent of each progeny, the parent's observed homozygosity, expected homozygosity for each progeny, average observed homozygosity for each progeny type, the range (min-max) of observed homozygosity, and the standard deviation (St.Dev).

	IBS				**IBD**			
Progeny Type	**S1**	**S2**	**S3-SB87-072**	**S3-SB87-073**	**S1**	**S2**	**S3-SB87-072**	**S3-SB87-073**
Count of individuals	37	48	41	48	37	48	41	48
Parent	IVP92-030-14	IVP92-030-14-3	SB87-072-P*	SB87-073-P*	IVP92-030-14	IVP92-030-14-3	SB87-072-P*	SB87-073-P*
Parent Observed Homozygosity	34.3%	56.5%	72.8 %	67.4 %	0%	33.8%	59.1 %	50.7 %
Expected Progeny Homozygosity	67.1%	78.2%	86.4 %	83.7 %	50%	66.9%	79.5 %	75.4 %
Average Observed homozygosity	61.4%	73.8%	83.4 %	82%	41.3%	60.3%	74.8 %	72.6 %
Range Observed homozygosity (Min-Max)	48.5-73.7%	65.2-82.2%	76.4-88.9 %	73.2-89.5 %	21.9-60%	47.2-72.9%	64.2-83%	59.3-84.2 %
St.Dev	8.6%	5.7%	4.3%	5.5%	12.9%	8.6%	6.4%	8.4%
Difference in Observed vs. Expected homozygosity	5.7%**	4.4%**	3.0%**	1.7%**	8.7%**	6.6%**	4.7%**	2.8%**
Average Observed homozygosity increase from parent	27.1%	17.3%	10.6 %	14.6 %	41.3%	26.5%	15.7 %	21.9 %

**The S2 parent plants for both S3 families were not available and were reconstructed from the offspring.*
***Statistical Significance (α = 0.05)*

The expected homozygosity, based on Mendelian inheritance, assumes a reduction in heterozygosity with each generation of selfing under random mating and absence of selection. However, these assumptions did not hold true for our breeding material, and we would theoretically anticipate higher homozygosity than predicted by the Mendelian model.

Consequently, the observed homozygosity was even lower than expected, indicating a greater deviation than hypothesised.

In conclusion, homozygosity increased progressively across the generations of selfing, with larger increases observed using the IBD approach compared to IBS. Significant deviations were found between observed and expected homozygosity levels, with the observed homozygosity being consistently lower than predicted by the Mendelian model. These discrepancies were more pronounced in the earlier generations and decreased with further inbreeding. Similar deviations from expected homozygosity have been reported in other potato studies, where observed heterozygosity levels remained higher than anticipated even after multiple generations of selfing (Song and Endelman 2023; van Lieshout et al. 2020).

Building on this analysis, we next explored RH, which refers to regions of the genome where heterozygosity persists despite the selfing process. Understanding these regions is crucial for elucidating genetic factors that may complicate the inbreeding process in potatoes.

Variation in homozygosity and heterozygosity across haplotags and targeted loci

In this study we aimed to identify alleles or genomic regions that resist homozygosity using haplotags in multiple loci which could indicate the presence of deleterious alleles. To achieve this we analysed the homozygosity and heterozygosity frequencies of 1,832 haplotags across all 442 diploid individuals in this study. Figure 5 presents bar plots of the frequencies of homozygosity (dosage = 2) and heterozygosity (dosage = 1) at each marker, with the haplotags ordered based on the number of clones exhibiting a specific dosage to facilitate clearer comparisons of allele dosage distributions. We observed that 652 haplotags were never in a homozygous state across all 442 clones, and 166 haplotags were homozygous in only one clone. Conversely, all other haplotags were observed in at least one clone in a heterozygous state, with 609 haplotags being heterozygous in between one and ten clones. The most frequent heterozygous haplotag, C3_8_00000000000, appeared in 351 individuals but was homozygous in only 17 individuals, while the second most common heterozygous haplotag, C3_8_00000011000, was present in 334 individuals and was never observed in a homozygous state.

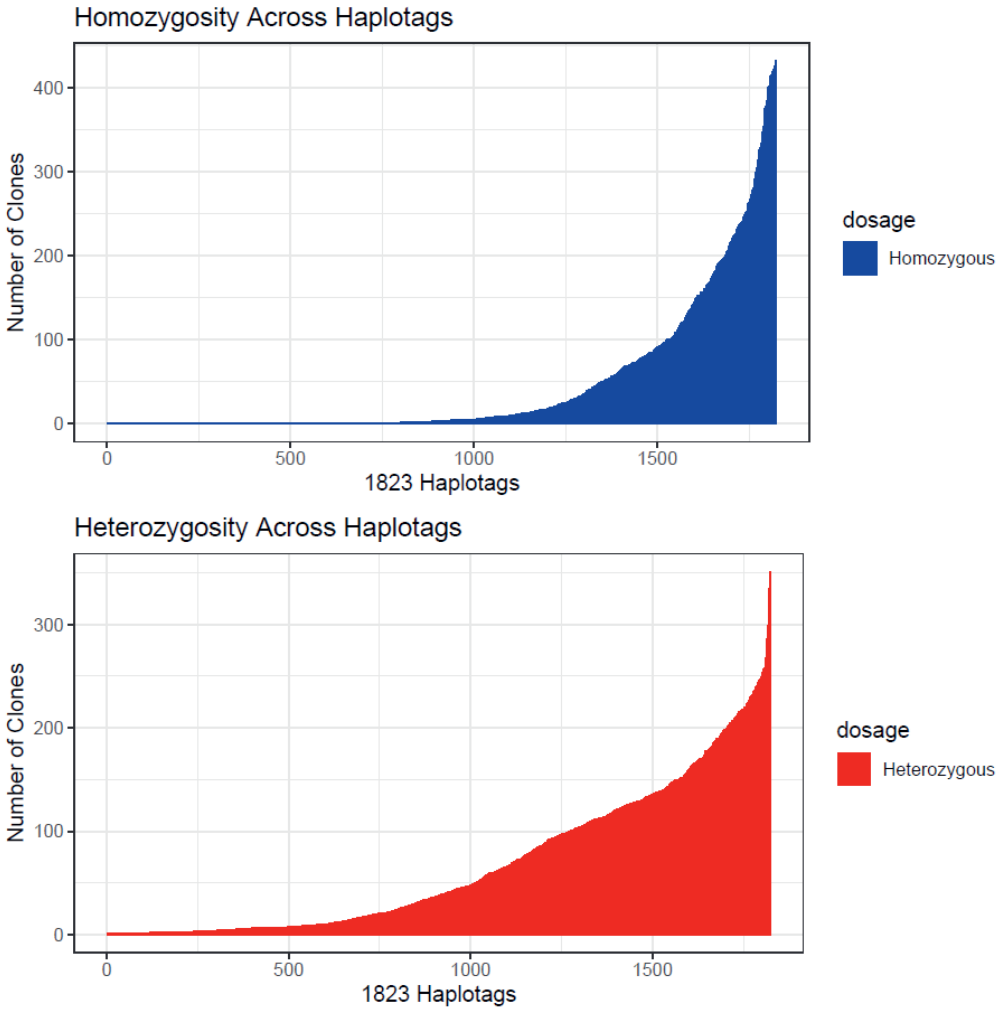


Figure 5. Frequency of (a) homozygosity and (b) heterozygosity across 1823 haplotags in 442 individuals. The x-axis represents markers sorted by the number of clones exhibiting a specific dosage (homozygous or heterozygous), and the y-axis shows the count of (a) homozygous or (b) heterozygous clones for each marker.

When examining the data from the point of view of individual loci rather than the component haplotags, the data exhibited variation in both homozygosity and heterozygosity. While all 333 loci could reach homozygosity (Figure 6), certain locisuch as C5_2, C3_8, and C12_10, showed a preference for the heterozygous state. Specifically, C5_2 was homozygous in only 2% of individuals, C3_8 in 6%, and C12_10 in 8%, suggesting that these regions may resist homozygosity.

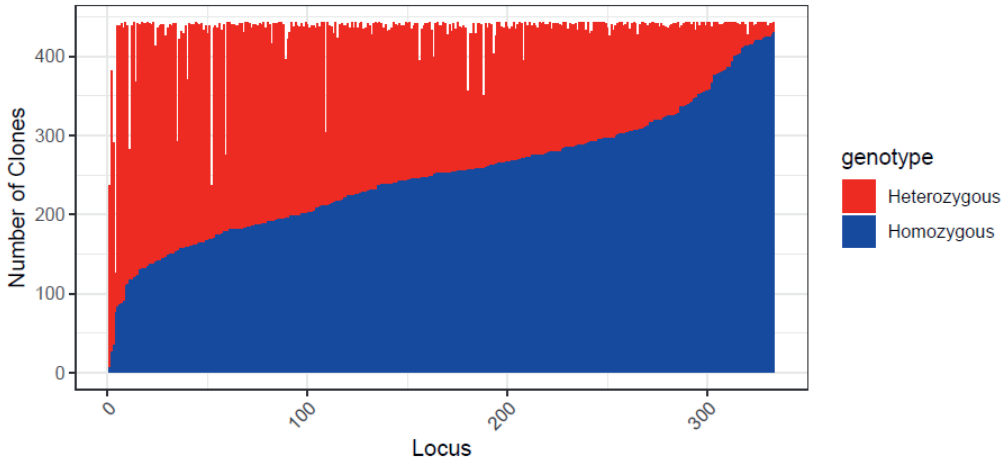


Figure 6. Frequency of homozygosity and heterozygosity across the 333 PotatoMASH loci in 442 individuals. The x-axis represents loci sorted by the number of clones with a specific dosage (homozygous or heterozygous), and the y-axis shows the count of individuals that were homozygous (blue) or heterozygous (red) at each locus.

In our previous GWAS study, a new QTL on chromosome 3 was identified as being associated with low yield and low tuber number in a panel that overlaps heavily with the material in this study (Vexler et al. 2024). Similarly, the locus C12_10, which was homozygous in only 6% of individuals, is situated near genes and QTLs linked to various performance and reproductive traits on chromosome 12. These traits include yield (McCord et al. 2011; Vexler et al. 2024), canopy vigour (Vexler et al. 2024), fruit set (Peterson et al. 2016), self-compatibility (Clot et al. 2020), and unreduced pollen production (Clot et al. 2024). By maintaining heterozygosity at these loci, the negative effects of deleterious alleles can be masked, thus potentially optimizing traits such as yield and fitness and preserving genetic diversity in the population. This observation underscores the critical role of selective forces in driving RH and preventing the expression of harmful alleles in plant breeding programs.

As outlined in the introduction, the current challenge in diploid hybrid breeding is that deleterious alleles are often genetically linked to beneficial alleles. This linkage complicates the inbreeding process, as efforts to fix beneficial alleles in homozygous states may inadvertently fix harmful alleles as well, leading to inbreeding depression. Loci that tend to remain heterozygous, like C3_8, provide insight into these regions where deleterious alleles are likely maintained in a heterozygous state due to genetic linkage with beneficial alleles. However, in the context of F1 hybrid breeding, this presents a complex issue. While these loci resist homozygosity, they do so because they are linked to essential traits for fitness and other desirable characteristics. As breeders endeavour to fix beneficial alleles through inbreeding, linked deleterious alleles may also be inherited, reducing overall fitness and yield. As an alternative, Fix-Res Breeding offers a potential solution by leveraging RH, maintaining these deleterious alleles in a heterozygous state to mask their effects while preserving the benefits of the beneficial alleles they are linked to. This strategy helps to

preserve heterozygosity for important traits, enhancing the breeding process and mitigating the risks of inbreeding depression and the fixation of harmful alleles.

Patterns of residual heterozygosity (RH) across potato chromosomes

In the context of plant breeding and population genetics, RH refers to the persistence of heterozygous regions in the genome despite multiple generations of selfing, which, under Mendelian expectations, should reduce heterozygosity by half with each generation. After five cycles of inbreeding, heterozygosity is typically expected to approach ~3%. However, in some genomic regions, heterozygosity persists across generations (Marand et al. 2019), suggesting resistance to fixation. A central question of this study is whether specific regions in diploid potato exhibit persistent heterozygosity or are more prone to fixation.

To explore this, we analysed 542 haplotags spanning 327 PotatoMASH loci in the founder IVP92-030-14 and its descendants. Homozygosity levels were calculated for each locus as the proportion of individuals that were homozygous within each progeny group and their respective parents. A detailed presentation of the IBS homozygosity levels for each progeny across all loci, including the 112 loci that were already homozygous in the founder, is provided in Suppl. File 7. Only loci that were heterozygous in the founder (n = 215) were included in Figure 7, which displays IBD-based homozygosity across the 12 chromosomes. Solid lines represent parental values; dashed lines show the average homozygosity across progeny.

Because Figure 7 includes only loci that were heterozygous in the founder, IVP92-030-14 (red) serves as a baseline, with homozygosity starting at zero across all regions. In its S1 progeny, average homozygosity levels ranged from 20% to 50% across loci. In the S1-derived parent IVP92-030-14-3 (purple), large segments of homozygosity had already formed—particularly on chromosomes 2, 10, and 12, and to a lesser extent on chromosomes 1, 6, 7, 8, 9, and 11. Further fixation was evident in the reconstructed parents of the S3 progenies. Notably, chromosome 11 reached complete homozygosity in S3-SB87-073. In contrast, chromosome 5 displayed persistent heterozygosity, with no locus reaching fixation in any generation.

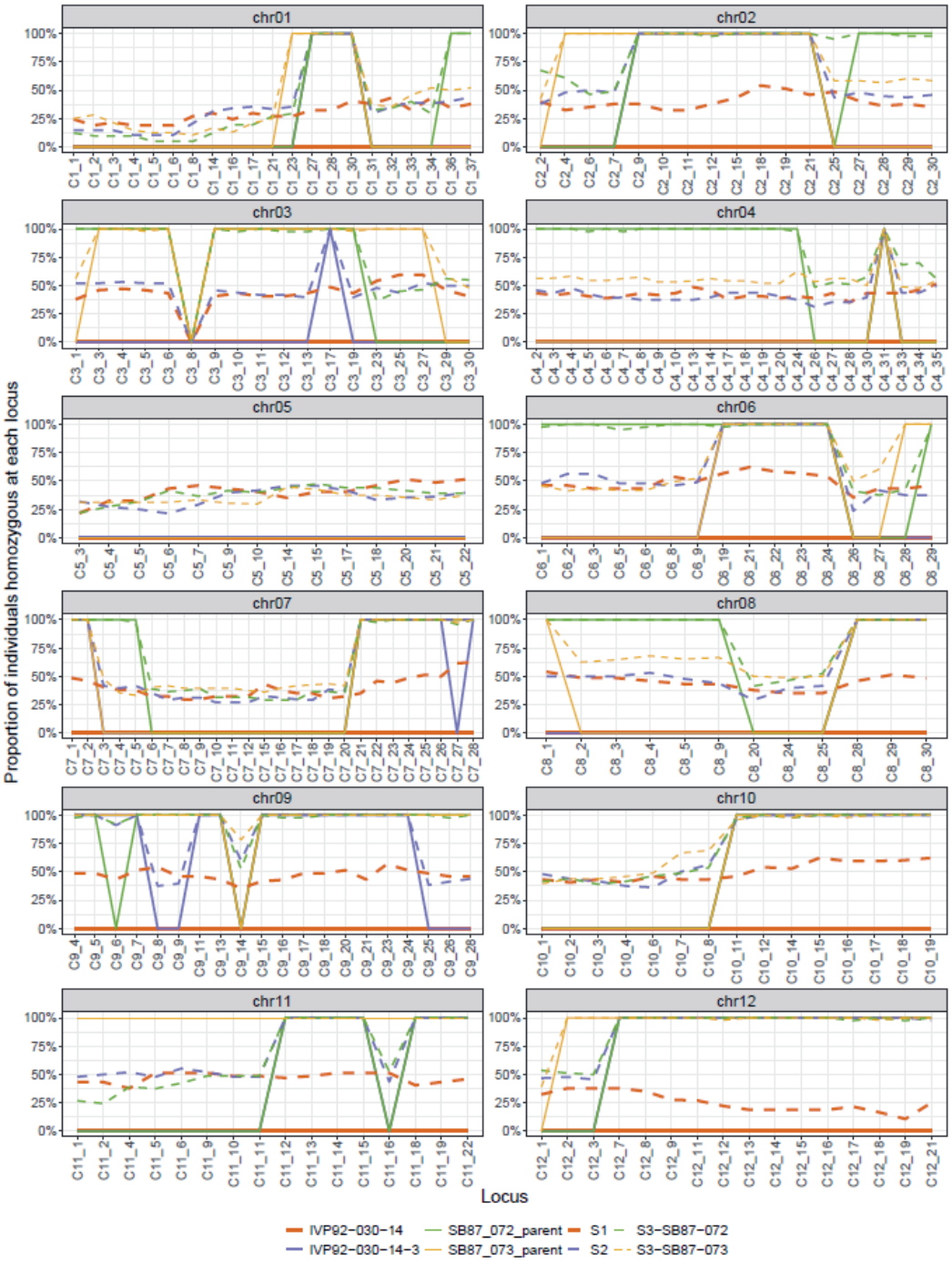


Figure 7. Proportion of individuals exhibiting homozygosity at each locus, based on haplotag data from the founder IVP92-030-14 and its descendants. Only loci that were heterozygous in the founder are shown. The y-axis indicates the proportion of individuals that are homozygous at a given locus; the x-axis shows loci

arranged by physical order within each chromosome. Each subplot represents one of the 12 potato chromosomes. Solid lines indicate parental homozygosity, while dashed lines show the average homozygosity across progeny. Parent and progeny generations are color-coded: IVP92-030-14 is the parent of the S1 (first selfed generation); IVP92-030-14-3 is the parent of the S2 (second generation); SB87-072-P and SB87-073-P are reconstructed parents of the S3 progenies (S3-SB87-072 and S3-SB87-073, respectively).

This observation is consistent with prior studies. It has been proposed that lethal or deleterious alleles might be maintained in repulsion with beneficial alleles in regions of reduced recombination (Jansky et al. 2014), and that increased recombination during sexual reproduction would be necessary to purge deleterious alleles (Leisner et al. 2018). However, Marand et al. (2019) found that RH in these regions was not due to a lack of recombination but instead driven by selective forces maintaining heterozygosity despite recombination. In the M6 genome (Marand et al. 2019), these heterozygous regions contained 6,878 genes, with a mean of 299 genes per block, and showed higher gene density. Marand et al. (2019) identified several selective forces contributing to the persistence of heterozygosity, including: (1) gametic selection, with genes linked to pollen development; (2) selection on floral tissue, with high gene expression in floral tissues; (3) epistatic selection, influencing the relationship between RH and traits like tuber number; and (4) overall yield and overall fitness, with RH being associated with enhanced yield and tuber production. These findings suggest that selective forces contribute to the maintenance of genetic diversity in these regions, which may confer advantages in reproductive success and overall fitness in potato populations.

Given the selective forces at play, we anticipate that regions resistant to homozygosity will vary between populations. In the M6 genome, heterozygous regions were primarily found on chromosomes 4, 7, 8, and 9, alongside shorter heterozygous blocks within homozygous regions on other chromosomes (Leisner et al. 2018; Marand et al. 2019). Our study also revealed a strong tendency toward RH in multiple regions. Specifically, region C3_8 exhibited consistent heterozygosity across all individuals in the experimental progenies, further highlighting the inability of this region to reach homozygosity through selfing. On chromosome 5, all 14 heterozygous loci in the founder remained segregating across the S1, S2, and S3 generations, although 25–50% of individuals were homozygous at these loci in each generation, none of the loci became fixed (Figure 7). This repeated pattern reflects a lack of transmission of homozygosity rather than an absence of homozygous individuals.

Given the well-documented effects of inbreeding depression in diploid potato, including reduced fertility and tuber production (Hosaka and Sanetomo 2020; Peterson et al. 2016; Phumichai and Hosaka 2006; Zhang et al. 2019; Wu et al. 2023), it is likely that individuals homozygous at key loci on chromosome 5 were not maintained due to poor performance. Since parental selection in this population was primarily based on phenotypic traits rather than homozygosity levels, such individuals may have been unintentionally excluded from further selfing. This would have led to the continued segregation observed in subsequent generations, illustrating how both natural and breeder-imposed selection can contribute to the persistence of heterozygosity in specific genomic regions.

Chromosome 5 is particularly noteworthy given previous reports linking this region to key traits, including maturity (Kloosterman et al. 2013), canopy growth (Vexler et al. 2024), unreduced gamete production (Clot et al. 2024), tuber number (Marand et al. 2019), and overall yield (Marand et al. 2019). Our findings suggest that, in the lineage we are observing, chromosome 5 harbours genes under selection that contribute to the persistence of heterozygosity.

Although homozygous individuals were observed at many loci on chromosome 5 in each generation, no locus reached complete fixation. One likely reason is that homozygous individuals were not consistently maintained or selected as parents. This may reflect a combination of natural and artificial selection: some highly homozygous individuals may have been non-viable due to inbreeding depression, for example, failing to flower or produce tubers, while others were likely excluded during routine selection by the breeder, who was not selecting based on homozygosity status but rather on agronomic performance and reproductive capacity. As a result, heterozygosity at these loci persisted across generations. This highlights how selection pressure, even when not explicitly genetic, can shape patterns of residual heterozygosity in breeding populations.

This case illustrates a broader challenge in diploid inbred development: balancing the need to fix beneficial alleles with the viability and fertility required for line advancement. Two general strategies exist in breeding: (1) population improvement followed by selfing to purge residual deleterious load, or (2) early selfing followed by intercrossing of viable partial inbreds. In either approach, generation-specific performance thresholds may be necessary, allowing lower performance, such as lower yield, in later selfed generations as homozygosity increases. These considerations are central to F1 hybrid breeding, which relies on fixing additive effects over time while restoring heterozygosity in the F1. In contrast, Fix-Res breeding offers a different route. Through the transmission of both homologous chromosomes via 2n gametes, Fix-Res preserves both additive and dominance components of the diploid genome in tetraploid offspring. This unique feature allows breeders to maintain beneficial heterozygosity in regions where fixation is unachievable or undesirable due to fitness costs, thereby reducing the pressure to develop fully homozygous inbred lines.

Conclusions

Tracking homozygosity is important in various diploid breeding strategies, including F1 hybrid breeding, and is particularly relevant for developing Fix-Res recurrent parents, where maintaining heterozygous loci contributes to optimizing specific traits. Inbreeding through selfing increases homozygosity; however, certain regions of heterozygosity may persist, potentially supporting genetic diversity and contributing to fitness and reproductive success in potato. Haplotag-based measurements provide a reliable method to assess homozygosity levels, offering valuable insights for advancement decisions without requiring parental genomic information. These measurements can also help identify loci that tend to remain homozygosity across generations in each population, whether due to selection, reduced

recombination, or fitness-related constraints, which is particularly relevant for the Fix-Res breeding approach.

Fix-Res breeding leverages RH by maintaining deleterious alleles in a heterozygous state, which masks their detrimental effects while preserving the benefits of linked favourable alleles. This approach optimizes clonal selection and prevents inbreeding depression and the fixation of harmful alleles.

Genotyping-by-Sequencing (GBS) and other high-density genomic methods can provide valuable insights into the evolving genetic structure of breeding panels under selection but can present some challenges. GBS approaches based on genome complexity with restriction enzymes are protected by patents and outsourcing costs are high, while development of array-based solutions requiring prior development can lead to ascertainment bias and are focussed on biallelic SNPs. Multiplex amplicon sequencing combined with read-backed haplotyping offers a potentially cost-effective genotyping tool for many applications, including the routine tracking of homozygosity in breeding populations. It delivers high accuracy by directly using haplotags that have been sequenced to a high coverage, which can provide more reliable homozygosity estimates than low-coverage GBS solutions.

PotatoMASH, a multiplex amplicon sequencing tool that enables read-backed haplotyping, provides a cost-effective and accessible solution for monitoring genetic parameters using haplotags. This system simplifies the assessment of genomic changes and selfing rates, eliminating the need for complex genotyping methods. By enabling precise measurements of homozygosity, its simplicity, flexibility, and ability to track genetic diversity make it an invaluable tool for supporting the accumulation of favourable alleles in Fix-Res breeding.

Designed to integrate seamlessly into potato breeding programs, PotatoMASH facilitates Marker-Assisted Selection (MAS) (Leyva-Pérez et al. 2022; Vexler et al. 2024), and routine homozygosity tracking. The genomic data produced every year with this approach for genomic-assisted breeding purposes effectively tracks genome-wide homozygosity changes and infers selfing rates, as demonstrated in this study, without the need for additional genotyping or further investment.

Acknowledgements

We thank the members of the public-private partnership "A new method for potato breeding: the 'Fixation-Restitution' approach" and SusCrop ERANET funded project "DIFFUGAT: Diploid Inbreds For Fixation, and Unreduced Gametes for Tetraploidy" (Averis Seeds B.V., Bejo Zaden B.V., Danespo A/S, Germicopa, Den Hartigh B.V., SaKa Pflanzenzucht GmbH & Co. KG, Meijer Potato, and Teagasc) for providing their support.

This research was carried out using the Teagasc high-performance computing cluster and storage systems, and the support of Dr. Paul Cormican is greatly appreciated.

Author contribution statements

H.v.E, D.M., V.P., R.H., and L.V., conceived and designed the study. V.P, C.E., and R.H. managed the breeding program, field trials and collected tissue samples. L.V. isolated DNA, constructed PotatoMASH libraries, analysed genomic data, preformed formal bioinformatics, statistical and homozygosity analysis, data curation, investigation, methodology and software and wrote the original draft. D.M., S.B., and H.v.E, advised on statistical analysis, methodology and visualization. M.d.l.O.L.-P. advised on bioinformatics processing. D.M., H.v.E., V.P., R.H., D.G., and R.G.F.V. obtained resources. D.M., H.v.E. and R.G.F.V. supervised the research. L.V., S.B., H.v.E., M.d.l.O.L.-P., V.P., R.G.F.V., and D.M. edited the manuscript. All authors reviewed and agreed to the published version of the manuscript.

Supplementary files

Suppl. File 1: Pedigree network of the diploid breeding program at PBR-WUR.

Suppl. File 2: PotatoMASH genotypic data for 442 diploid inbred clones in the study.

Suppl. File 3: IBS homozygosity for the germplasm from the Wageningen diploid breeding program.

Suppl. File 4: IBS and IBD homozygosity 67 breeding clones with parents.

Suppl. File 5: IBS and IBD homozygosity experimental progenies.

Suppl. File 6: Additional visualizations and data summary for the manuscript.

Suppl. File 7: Variation in homozygosity levels with all 327 PotatoMASH loci (including heterozygous loci).

Chapter 5

PotatoMASH is a cost-effective marker system for Genomic Prediction in potato based on short-read haplotypes

Authors

Lea Vexler[1,2,3], Agnieszka Konkolewska[1,4], Stephen Byrne[1], Tom Ruttink[5,6], Maria de la O Leyva- Pérez [1], Jie Kang[7,8,9], Denis Griffin[1], Richard G.F. Visser[2], Herman J. van Eck[2], Dan Milbourne[1]*

Affiliations

[1]Teagasc, Crop Science Department, Oak Park, R93 XE12 Carlow, Ireland

[2]Plant Breeding, Wageningen University & Research, P.O. Box 386, 6700 AJ Wageningen, The Netherlands

[3]Graduate School Experimental Plant Sciences, Wageningen University & Research, Wageningen, The Netherlands

[4]Insight SFI Research Centre for Data Analytics, School of Computer Science, University College Dublin, Dublin, Ireland

[5]Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Plant Sciences Unit, Caritasstraat 39, Melle 9090, Belgium

[6]Department of Plant Biotechnology and Bioinformatics, Faculty of Sciences, Ghent University, Technologiepark 71, Ghent 9052, Belgium

[7]Charles Perkins Centre, The University of Sydney, Camperdown, NSW, Australia

[9]Sydney Precision Data Science Centre, The University of Sydney, Camperdown, NSW, Australia

[9]School of Mathematics and Statistics, The University of Sydney, Camperdown, NSW, Australia

Submitted to Theoretical and Applied Genetics (April 2025)

Abstract:

Genomic Prediction (GP) supports plant breeding by accelerating genetic improvement; however, the high cost associated with dense genotyping platforms restricts their use in routine breeding. This study evaluates the efficacy of PotatoMASH, a cost-effective amplicon-sequencing platform generating SNPs and short-read multi-allelic haplotypes (haplotags), for GP in potato. In a tetraploid population, we assessed the prediction accuracy (PA) of PotatoMASH data in comparison to GBS data for the complex trait fry colour. Utilising only 2,236 SNPs and 2,000–3,390 haplotags from 339 amplicon loci (versus 43.6k SNPs from GBS), PA was moderately reduced by 14% using SNPs and only 9% using haplotags. In a diploid panel, PotatoMASH was applied for GP of 23 agronomic, quality, and morphological traits. Both marker types achieved medium to high PA across all traits (0.29–0.81), with small performance differences: haplotags outperformed SNPs in 11 traits, while SNPs performed better in six. Additionally, we evaluated two haplotyping methods implemented in the SMAP software: haplotype-sites, which combines variants at pre-called SNPs, and haplotype-window, which extracts full read sequences from defined genomic windows. Haplotags derived from haplotype-window outperformed those from haplotype-sites in six traits. This research demonstrates that PotatoMASH, which facilitates the concurrent detection of both SNPs and haplotypes at a reduced cost, represents a scalable alternative to GBS. It provides a versatile and economical genotyping solution suitable for integrated pipelines that combine marker-assisted selection (MAS) and GP in potato breeding.

Introduction

Potato breeding is a protracted process, as most potato breeders employ conventional phenotypic selection for tetraploid potatoes. The outbreeding nature and tetraploidy of modern cultivated potatoes present significant challenges in achieving enhanced genetic gains. The capacity to rapidly utilise recurrent selection to fix favourable alleles and eliminate deleterious alleles across multiple cycles of selection is constrained.

In the post-genome era, the utilisation of genomic tools to achieve higher and faster genetic gain in the breeding process has been extensively investigated. Many plant breeding programmes now routinely incorporate Marker-Assisted Selection (MAS) select favourable alleles for various traits. MAS is most effective for monogenic traits, or those with a few QTL with large effects, but it is less effective in complex traits where the genetic variation results from a larger number of loci of small effects (Heffner et al. 2009).

Genomic selection (GS) is a form of MAS in which genetic markers spread across the whole genome are used to predict breeding values of individuals for the purpose of ranking selection candidates in practical breeding (Goddard and Hayes 2007; Heffner et al. 2009; Meuwissen et al. 2001). By incorporating genome-wide markers in the Genomic Prediction (GP) model, GS utilises a greater proportion of the variation due to QTL with small to moderate effects (Goddard and Hayes 2007; Heffner et al. 2009).

As many key agronomic traits in potato are highly polygenic and environmentally sensitive, GS offers a particularly robust framework for their improvement. Such traits are typically governed by numerous small-effect loci and are further complicated by epistasis, regulatory interactions, and extended linkage disequilibrium (LD), all of which obscure the identification of causal variants (Hill et al. 2008; Kloosterman et al. 2010; Stich and Gebhardt 2011). For instance, yield is governed by complex gene networks influencing root architecture, shoot development, tuberization, and carbohydrate metabolism (Kacheyo et al. 2021; Qu et al. 2024; Navarro et al. 2011; Kloosterman et al. 2008; Ewing and Struik 1992). Similarly, tuber quality traits such as dry matter content, chipping colour, and dormancy are affected by both genetic variation and environmental factors, including storage temperature, physiological maturity, and stress exposure (Hu et al. 2023; Leonel et al. 2017; Stark et al. 2020; Zhou et al. 2017). In particular, regions of high LD can lead to the co-localization of multiple QTLs, making it difficult to resolve the genetic effects of individual traits. This phenomenon has been widely reported in QTL studies for complex traits in potato (Acharjee et al. 2018; Kloosterman et al. 2013; Sharma et al. 2018; Vexler et al. 2024). Although major and minor QTLs have been identified for many traits, the associated markers often lack resolution and transferability. In many cases, markers detected in one population are physically distant from the causal polymorphisms in another, and are difficult to apply in breeding material with different genetic origins (Kloosterman et al. 2010). These limitations reduce the practical utility of MAS for improving complex traits.

The potential of GS to address certain complexities in the breeding process has attracted significant interest within the potato breeding community. Several studies have evaluated the efficacy of various statistical methods and machine learning approaches for the accurate prediction of Genomic Estimated Breeding Values (GEBVs) across a range of agronomic performance traits (Aalborg and Nielsen 2024; Aalborg et al. 2024; Adams et al. 2023; Byrne et al. 2020; Ortiz et al. 2022; Ortiz et al. 2023; Pandey et al. 2023; Selga et al. 2021b; Slater et al. 2016; Stich and Van Inghelandt 2018; Sverrisdóttir et al. 2017; Sverrisdóttir et al. 2018; Wilson et al. 2021).

For successful implementation of GS, it has been suggested that high-density Single Nucleotide Polymorphism (SNP) panels are necessary to capture genetic variation across the entire genome and to ensure that all QTL are in LD with at least one marker, allowing them to capture a significant portion of the genetic variance (Heffner et al. 2009; Slater et al. 2016). In fact, most GS studies in potatoes rely on genotyping-by-sequencing (GBS) data containing tens to hundreds of thousands of SNPs. For example, the 186k SNPs (Sverrisdóttir et al. 2017), 46k SNPs (Byrne et al. 2020), 39k SNPs (Wilson et al. 2021), or SNP array such as the 8.3k SNPs from the SolCAP potato genotyping array (Stich and Van Inghelandt 2018).

The option of utilising reduced marker sets for prediction in potatoes has been investigated in recent studies, and various strategies have been employed to reduce the number of SNPs. These approaches include LD pruning based on the estimation of LD between adjacent SNPs (Selga et al. 2021b), employing selected markers from a Genome-Wide Association Study (GWAS) (Byrne et al. 2020), and utilising feature selection approaches to identify subsets of SNP variants (Aalborg et al. 2024). In all cases, sufficient predictive ability was achieved, albeit lower than that obtained with genome-wide markers. Consequently, smaller (and therefore cheaper) marker sets can be employed for GP in potatoes. The ongoing challenge lies in reducing the number of markers to minimize the costs of genotyping and data processing/storage while maximizing the allelic information to support the performance of prediction models.

One strategy for improving GP models is the use of multi-allelic haplotypes instead of bi-allelic SNPs. The initial hypothesis proposed that the incorporation of haplotype alleles in prediction models would result in increased accuracy compared to individual SNPs, owing to the presumption that haplotype alleles are likely in stronger LD with QTL alleles than SNP alleles (Calus et al. 2008; Hess et al. 2017; Meuwissen et al. 2014). This has already been demonstrated by many studies in animal breeding, which showed that haplotype alleles can enhance the accuracy of GP compared to SNPs (Araujo et al. 2023; Cuyabano et al. 2014; Hess et al. 2017; Li et al. 2022; Won et al. 2020). Although research on using haplotype alleles in plant breeding is more limited, several studies have investigated aggregating adjacent SNPs into haplotype blocks, typically containing 2 to 20 SNPs, in crops such as wheat (Sallam et al. 2020), *Eucalyptus globulus* (Ballesta et al. 2019), rice, and maize (Jiang et al. 2018). A more recent study in wheat explored long-range haplotypes with blocks up to

100 SNPs (Difabachew et al. 2023), and a study on ryegrass evaluated the predictive performance of SNP-based haplotypes from GBS-simulated data (Kang 2023). These studies consistently showed improved prediction accuracy (PA) for most traits when haplotype blocks were used instead of individual SNPs.

In animal breeding, haplotype construction typically involves large SNP arrays and either statistical haplotyping or phasing SNPs based on known parental inheritance. These methods can result in long haplotypes, sometimes spanning entire chromosomes, and are implemented with software such as BEAGLE (Browning and Browning 2009; Browning and Browning 2007). However, challenges arise when working with rare alleles, polyploids or populations with high genetic diversity, where long haplotypes may be inaccurate, and rare variants may be misinterpreted (Garrison and Marth 2016). In plant breeding, haplotype blocks are often constructed by ordering SNP markers based on consensus map positions (Sallam et al. 2020; Jiang et al. 2018) or by using LD-based statistical methods (Difabachew et al. 2023), with tools like Haploview (Barrett et al. 2005). However, these methods also encounter difficulties, particularly when working with polyploids, or when marker density is low or in low-depth sequencing scenarios.

Here, we use Stack Mapping Anchor Points (SMAP) (Schaumont et al. 2022) for read-backed haplotyping using the sequence information within individual NGS reads. SMAP performs accurate read processing and analyses mapping distributions across sample sets. Its key advantage is the ability to reconstruct actual haplotypes directly from short sequencing reads (Schaumont et al. 2022). This approach circumvents several challenges associated with traditional methods, especially in low-coverage sequencing contexts and irrespective of polyploidy, offering a more robust way to infer haplotypes directly from sequencing data. SMAP includes two haplotyping modules: *haplotype-sites*, which joins pre-called SNPs into a haplotag using read backed phasing, and *haplotype-window*, a newer module that works without prior variant calls by extracting the full-length DNA sequence of reads that are mapped to a particular window in the reference sequence and contain two flanking locus-specific border sequences. The latter has the potential to detect a broader range of variants, including SNPs, indels, and any combination thereof, and improve haplotype resolution. A detailed description of both modules is provided in the Materials and Methods section.

A major limitation in applying GS to potato breeding is the high cost of genotyping. Despite the decreasing costs of next-generation sequencing (NGS), genotyping thousands of seedlings annually for GS remains expensive. GBS is patented (http://www.google.com/patents/US8815512) and requires licensing and further royalty fees for use in commercial cultivar development. As a result, the cost of GBS, when commercially sourced, ranges from 50 to 100 euros per sample, depending on the specific characteristics of the sample. This significant expense presents a challenge for widespread adoption of GS in potato breeding, highlighting the need for more cost-effective alternatives without compromising the accuracy of GPs.

In a previous study (Leyva-Pérez et al. 2022), we developed a marker system called PotatoMASH with the specific objective of exploring the potential of low-cost, genome-wide genotyping for application in potato breeding and genetics. PotatoMASH, with an approximate cost of 4-5 euros per sample, surveys 339 loci using a multiplex amplicon sequencing approach followed by deep NGS sequencing (2x150 bp Illumina sequencing). Although the system is low-density in terms of loci number, it was specifically designed to achieve genome-wide coverage by evenly spacing markers approximately every 1 cM across euchromatic regions, while keeping genotyping costs low. Due to the high SNP density in potato germplasm, PotatoMASH yields more than 2000 SNPs in diverse panels (Leyva-Pérez et al. 2022; Vexler et al. 2024), and over 800 SNPs in bi-parental populations (Clot et al. 2024). Additional tools can be utilised for read-backed phasing (Schaumont et al. 2022) to generate short haplotypes (165-180 bp) that represent the allelic diversity at each locus.

To assess the effectiveness of short-read haplotypes from PotatoMASH for QTL detection, Leyva-Pérez et al. (2022) genotyped a tetraploid population using PotatoMASH, generating two marker sets: a SNP set (2279 SNPs) and a set of short-read haplotype alleles (hereafter referred to as haplotags) derived from these SNPs (2000 haplotags). While the PotatoMASH SNP set failed to detect any QTL, the haplotag set successfully identified the same QTL associated with fry colour that had been previously detected in a GWAS using 43.6k GBS-derived SNP markers (Byrne et al. 2020). This finding suggested that haplotags may offer better discriminatory power than SNPs for QTL detection in GWAS.

With this expectation, in a subsequent study (Vexler et al. 2024), we used PotatoMASH to generate a SNP marker set and then applied SMAP *haplotype-sites* software to derive a haplotag marker set. These marker sets were subsequently used to conduct a GWAS on a diploid panel for a range of traits. We detected 37 unique QTL across both marker types. Interestingly, the haplotags did not consistently outperform the bi-allelic SNPs in QTL detection for all the traits, but instead, about 30% of QTL were identified using the SNP set but not the haplotag set. Further investigation of this phenomenon revealed that for QTL detected only with haplotag data, the significant haplotags often contained individual SNPs that were also present in other non-significant haplotags. Conversely, for QTL detected only with SNP data, the significant SNPs were dispersed across multiple haplotags, each with a lower frequency in the population than the SNP itself, which reduced the statistical power to detect associations. These findings led to the conclusion that both SNPs and SNP-based haplotags should be used in parallel to maximize QTL detection power (Vexler et al. 2024). Based on the differential performance of SNPs and haplotags in GWAS, we decided to explore their relative effectiveness for GP in the populations described above.

The objectives of this study were (1) to demonstrate that PotatoMASH, as a genome-wide, low-density marker set, can achieve comparable accuracy in GP to high-density marker sets like GBS; (2) to explore ways to increase genomic information generated by PotatoMASH by comparing two haplotype-calling approaches: the previously used *haplotype-sites* and a

newly introduced *haplotype-window* approach using SMAP; (3) to test the prediction accuracy (PA) in a diploid potato breeding population to assess PotatoMASH's effectiveness across a broad range of traits.

Material and methods

Short-read haplotype allele (haplotags) calling with SMAP

We utilised SMAP (Schaumont et al. 2022) for read-backed haplotyping to call short-read haplotype alleles (haplotags) in our study. SMAP overcomes some of the limitations of traditional haplotyping tools, which often have restricted applications, such as being applicable only for diploid individuals, requiring fixed haplotype block lengths, or relying on prior knowledge of genetic relationships or variants.

SMAP works by delineating loci using customized start and end points per haplotag, e.g. PotatoMASH multiplex primer binding sites, and offers two complementary modules for haplotype calling: SMAP *haplotype-sites* and SMAP *haplotype-window*. The two approaches are illustrated in Fig. 1.

SMAP *haplotype-sites*

The first module, SMAP *haplotype-sites* (see online manual at https://ngs-smap.readthedocs.io/en/latest/sites/index.html), phases genotype calls at pre-defined 'sites' (SNPs). Per read mapped to a particular locus, it reconstructs short haplotype alleles (haplotags) based on the polymorphisms it detects at the positions of these SNPs. The input for this module includes indexed BAM files with mapped reads, a custom BED file with locus start and end positions (in our case, 339 PotatoMASH loci), and a VCF file with SNP positions (called ‘sites’). The process involves extracting polymorphic sites from each read, encoding the presence of the reference allele as "0", alternative allele as "1", absent mapping as "." or a gap "-", and creating a compressed haplotype string (e.g. 000110-10) for each read. The final haplotag nomenclature we used includes the PotatoMASH locus name (e.g., C1_1) and the haplotype string given by SMAP *haplotype-sites* (e.g., C1_1_000110-10). In previous studies, we have successfully used this approach for SNP-based haplotag calling both for tetraploid (Leyva-Pérez et al. 2022) and diploid (Vexler et al. 2024) potato panels. The full bioinformatics pipeline of PotatoMASH and SMAP *haplotype-sites* and the intermediate files needed are also available at https://doi.org/10.6084/m9.figshare.c.6926560.

SMAP *haplotype-window*

The second module, SMAP *haplotype-window* (see online manual at https://ngs-smap.readthedocs.io/en/latest/window/index.html), operates without prior knowledge of polymorphisms. It defines a "window" for each locus using two locus-specific border sequences. Coordinates for these borders are provided as a custom GFF file (for PotatoMASH regions in DM_v6.1 is provided in Suppl. File 1). The module identifies all reads mapped to the locus, retrieves the corresponding sequences from the original FASTQ

files, and trims the border sequences to retain the remaining DNA sequence (which forms a unique haplotype allele (haplotag)). This method allows for the detection of haplotags without the need for pre-existing SNP calls and can capture a broader range of genetic variants, particularly indels. In this study, we also applied SMAP *haplotype-window* to generate haplotags with different parameters for tetraploid and diploid panels:

SMAP haplotype-window v5.0.1 was run with the following parameters for tetraploids: –discrete_calls dosage –frequency_interval_bounds 12.5 12.5 37.5 37.5 62.5 62.5 87.5 87.5 –dosage_filter 4 –min_read_count 20 –min_haplotype_frequency 10 –min_distinct_haplotypes 0.

For diploids: –discrete_calls dosage –frequency_interval_bounds 10 10 90 90 –dosage_filter 2 –min_read_count 10 –min_haplotype_frequency 10 –min_distinct_haplotypes 0.

With *haplotype-window*, haplotype nomenclature that is given by the software SMAP is the actual unique read sequence. The final haplotag name is the PotatoMASH locus name followed by the unique read sequence (e.g., C1_1_ATTGGGTTCCACACTTTTGACTATGCGAGGCACTTCCTCTCGTGTTGCAGTCGGATGTTAGGCATTTCTTATGAGTCGAAAAGGGGTTACATAGGCCTCGAGTACTATGGTAGGACTGTAAGTATTAAA).

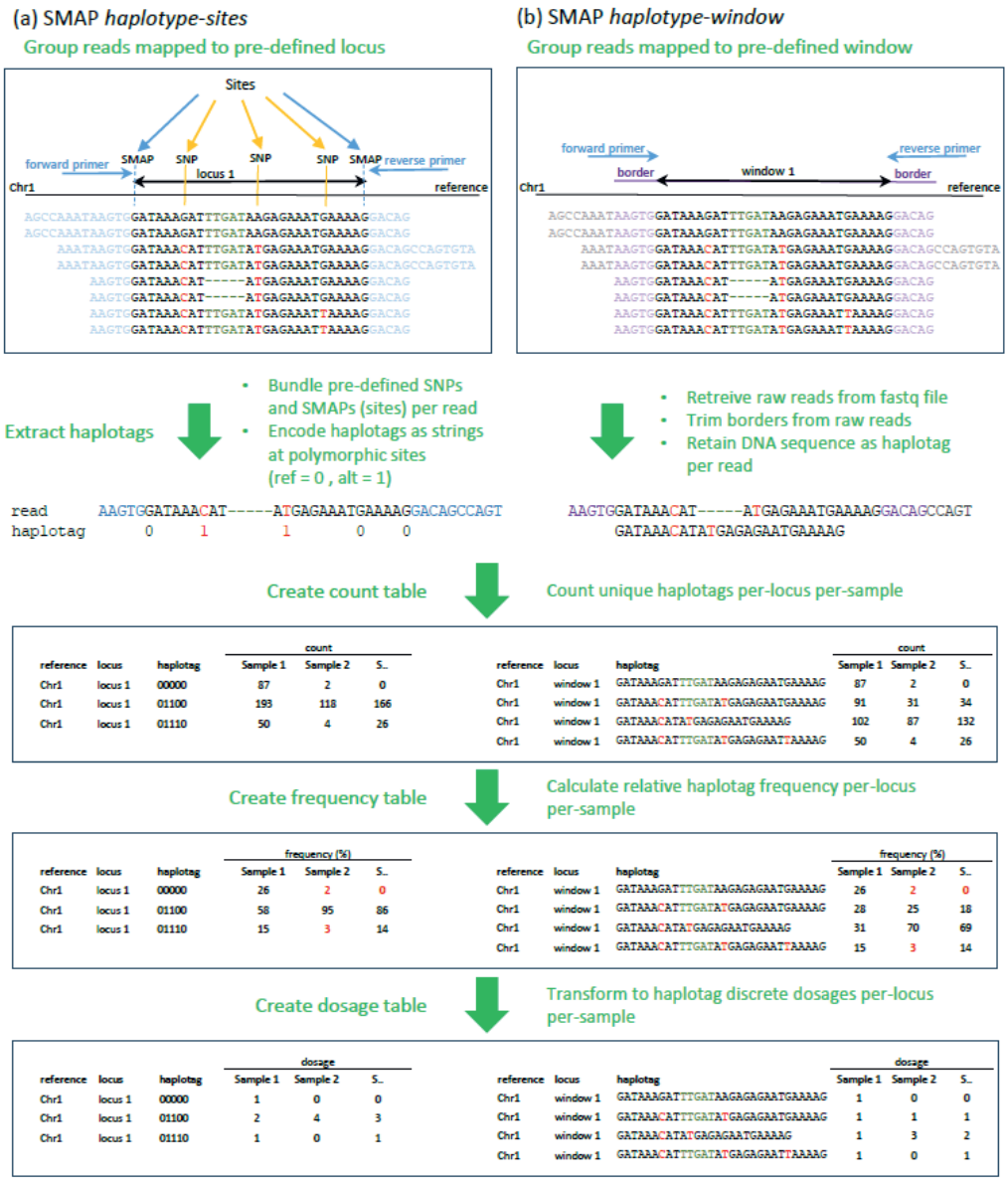


Fig. 1. *Comparison of SMAP haplotype-sites and haplotype-window modules for haplotags construction from read alignments.* (*a*) Haplotype-sites *groups sequencing reads mapped to pre-defined loci and reconstructs multi-allelic haplotags by bundling pre-defined SNPs and SMAPs (Stack Mapping Anchor Points). Each read is transformed into a coded string ("0" for reference, "1" for alternate alleles and "." or "-" for missing or gapped positions), representing the allelic pattern at the locus. This approach requires prior knowledge of variant positions and encodes haplotags and does not capture indels, null alleles or triSNPs. (b) Haplotype-window identifies reads mapped between user-defined genomic windows using pre-defined border sequences (e.g., primer sites), trims the borders, and retains the full DNA sequence between them as a unique haplotag. It operates without requiring prior variant calls and captures a broader range of variation, including indels, enabling more haplotypes to be distinguished per locus. This example illustrates haplotags calling in a*

tetraploid context using the same sequencing data: haplotype-window resolves four haplotypes compared to three identified by haplotype-sites, due to additional sequence variation outside the pre-defined SNPs. Both modules generate tables of haplotag counts, allele frequencies, and inferred allele dosages per-locus per-sample, where numbers in red indicate the frequency of sequencing errors before their removal. Figure adapted from the official SMAP documentation (https://ngs-smap.readthedocs.io) to illustrate the logic and data structure of both approaches.

Comparison of Haplotype Calling Modules

Both modules generate an integrated haplotype call table that lists all unique haplotag counts per locus and sample. They apply the same functionality for haplotag frequency filtering, discrete genotype calling, and quality controls for loci and samples. Additionally, both modules create summary statistics, which are represented in tabular and graphical formats. Since both modules extract haplotags in a different way, while using the same mapped reads as input, we anticipate subtle differences in the final haplotype calls that can impact downstream analysis and data archiving. This provides a strong motivation for comparing the performance of both haplotype calling modules.

Firstly, SMAP *haplotype-window* does not rely on prior SNP calling, which reduces the risk of false negatives that could hinder downstream haplotype detection (Veeckman et al. 2019; Garrison and Marth 2016). Second, as *haplotype-window* is not restricted to pre-defined SNPs, it has the ability to capture a broader range of genetic variance, including indels, leading to the identification of a greater number of haplotags. Finally, *haplotype-window* ensures consistent naming of haplotypes across datasets, eliminating the need for complete re-analysis of genotyped germplasm upon identification of novel SNPs, further streamlining data maintenance and comparison.

Plant materials and phenotypic evaluation

Tetraploid panel

We used a set of 607 tetraploid potato clones, known as the FRY population. GP for fry colour had previously been conducted on this population using approximately 43.6k GBS-derived SNP markers (Byrne et al. 2020). The phenotypic data for fry colour 'off-the-field' (OTF) were generated in the Teagasc/IPM breeding programme in 2015-2017 and included 237 individuals in 2015, 73 in 2016, and 297 in 2017 (Byrne et al. 2020). This panel was subsequently utilised to test PotatoMASH (Leyva-Pérez et al. 2022). The sequence data produced from PotatoMASH can be processed in one of two ways; (i) to identify and genotype at SNP positions, and (ii) to use read-backed phasing to generate short haplotags. In the case of the FRY population, a set of 2279 filtered SNPs were identified and these SNP positions were further utilised to identify a set of 2000 haplotags using the SMAP *haplotype-sites* module ((Leyva-Pérez et al. 2022). Here, we use both these data sets for GP. In addition, we used SMAP *haplotype-window* as an alternative approach to identify sequence-based haplotags.

Diploid panel

A panel of 558 elite diploid breeding clones was provided by a consortium comprising commercial breeders and research institutes. The panel represents clones from diploid breeding programmes, where commercially relevant traits are combined with traits significant for diploid breeding such as fertility, self-compatibility and unreduced gamete production. Contributions were made by Meijer Potato (The Netherlands); Plant Breeding, Wageningen University & Research (The Netherlands); Danespo A/S (Denmark); SaKa Pflanzenzucht GmbH & Co. KG (Germany); Germicopa Breeding (France) and Averis Seeds B.V. (The Netherlands).

A comprehensive description of the panel, experimental design, genotyping, variant calling for the SNPs, SNPs-based haplotags, phenotyping methods and the statistical analysis are described in Vexler et al. (2024). In short, the experimental design followed an augmented design with replicated checks across the six locations. The material was evaluated over three years (2019-2021) using standardized set of protocols for scoring 23 morphological, agronomic and quality traits (Table 1). Check varieties were included in the estimation of phenotypic means for all traits across years and locations. Best Linear Unbiased Estimators (BLUEs) for the three year observations were calculated using the lme4 R package (Bates et al. 2015) and the *lmer* function. Least square means were calculated for the BLUEs with the R package emmeans (Lenth et al. 2021), and served as the final phenotypic data for GP. Broad-sense heritabilities (H_2) were calculated for each sub-population separately. The panel was genotyped with PotatoMASH resulting in 2730 filtered SNPs and 2955 haplotags identified by SMAP *haplotype-sites* with these SNPs (Vexler et al. 2024).

We utilised this panel, the SNPs set, and the haplotags set to perform GP for the 23 traits. We have also employed the sequenced PotatoMASH reads of this panel to call haplotags with SMAP *haplotype-window*, as detailed above.

Table 1. Phenotypic scoring and measurements of the 23 traits evaluated in the diploid panel (Vexler et al. 2024)

Trait	**Scale**
Yield	In kg per plant, fresh weight at harvest
Canopy stage 1 (6 weeks after planting)	1 = plants have not yet emerged to 9 = largest canopy in the trial
Canopy stage 2 (10 weeks after planting)	1 = plants have not yet emerged to 9 = largest canopy in the trial
Tuber Length	Tubers per meter count was used with correction table
Total Tuber Number	Count of tubers
Tuber Shape	1 = very round, 2 = round, 3 = round-oval, 4 = round-oval to oval, 5 = oval, 6 = oval to long-oval, 7 = long-oval, 8 = long, 9 = very long
Yellow Skin Colour	1 = white, 2 = cream, 3 = light yellow, 4 = yellow, 5 = dark yellow, 6 = brown
Yellow Flesh Colour	1 = clear white, 2 = white, 3 = cream, 4 = light yellow, 5 = yellow, 6 = dark yellow, 7 = very dark yellow
Eye Depth	1 = very deep to 9 = very shallow
Presentability of Tubers	1 = very bad to 9 = very good
Skin Smoothness	1 = rough to 9 = very smooth
Skin Brightness	1 = dull to 9 = clear
Sensitivity to Common Scab	1 = heavy symptoms to 9 = no symptoms
Enzymatic Browning	1 = ink black, 2 = uniformly black, 3 = discolouration to black, 4 = darkening of red and grey discolouration, 5 = bright red and dark grey discolouration, 6 = start of red/grey discolouration, 7 = clear start of discolouration, 8 = very slight discolouration, 9 = no discolouration
Cooking Type	2 = very floury, loose boiling, sloughing, 4 = floury, crumbly and fairly loose, 6 = slightly floury and fairly firm, 8 = not floury, firm cooking, 9 = extreme firmness
After-cooking blackening	1 = very dark to 9 = pure colour (no darkening at all)
Chipping Colour 1st time point stored at 8°C	1 = very dark to 9 = pure colour (no darkening at all)
Chipping Colour 2nd time point stored at 8°C	1 = very dark to 9 = pure colour (no darkening at all)
Chipping Colour 2nd time point stored at 4°C	1 = very dark to 9 = pure colour (no darkening at all)
Dry Matter Content	% relative to fresh weight
Sprout Dormancy	1 = heavy sprouting (early) to 9 = no sprouting
Tuber Regularity	1 = bad to 9 = good
Maturity	1 = plants still green and flowering to 9 = plants reached senescence

Genomic Prediction

Genomic prediction was performed using three marker sets: SNPs, SNP-based haplotags generated by SMAP *haplotype-sites*, and sequence-based haplotags generated by SMAP *haplotype-window*. For SNPs, the dosage was defined as 0, 1, or 2, where 0 represents homozygosity for the reference allele, 1 represents heterozygosity, and 2 represents homozygosity for the alternative allele. For the haplotags, the discrete dosage of each haplotype was scored, with 0 indicating the haplotype is absent, 1, 2, 3, or 4 as discrete allele dosage for tetraploid individuals, and 1 or 2 for diploid individuals. Haplotags were treated as "pseudo-SNPs," with each individual haplotype effectively rated as a bi-allelic presence-absence marker. This marker dosage scale was incorporated into the Bayesian models. For ridge-regression Best Linear Unbiased Prediction (rrBLUP) and Genomic Best Linear Unbiased Prediction (GBLUP), the dosage was encoded as -1, 0, and 1, following the requirements of the rrBLUP R package (Endelman 2011), where 0 represents heterozygous states, and -1 and 1 represent homozygous states.

Allele frequencies for each marker were calculated by taking the mean genotype value across the population, based on the actual dosage and normalized to the maximum possible genotype value (4 for tetraploids and 2 for diploids). Minor allele frequencies (MAF) were then determined as the smaller value between the allele frequency and its complement.

Genomic Prediction in the tetraploid population

We applied the same GP framework and the same four statistical algorithms for GP used by Byrne et al. (2020): ridge regression best linear unbiased predictor (rrBLUP R Package) (Endelman 2011), Bayes A (Meuwissen et al. 2001), Bayesian Lasso (Park and Casella 2008) and Random Forest (Liaw and Wiener 2002). The two Bayesian approaches were implemented in the R package BGLR (Pérez and de los Campos 2014) with the following parameters: number of iterations = 5000, burn-in =500 and thinning = 5. Random forest was implemented with the R package Random Forest (setting the number of variables at each split to 1/3 of the total variables and using a terminal node size of five and minimum of 500 trees per forest). PA was calculated as the Pearson correlation coefficient between observed and predicted OTF fry colour Best Linear Unbiased Predictions (BLUPs) values. GP models were developed for each year separately (2015–2017) and evaluated in other years, as in the original study (Byrne et al. 2020).

Genomic Prediction in the diploid panel

Genomic data were centred around zero, and mean imputation was used to replace missing values. Six models were evaluated: GBLUP with the *kin.blup* function, rrBLUP with the *mixed.solve* function from the rrBLUP package (Endelman 2011), and Bayesian models: BayesA, BayesB, BayesC, BRR and BL from the BGLR package (Pérez and de los Campos 2014). Default options were used unless stated otherwise.

We employed the following approach for GP: K-fold cross-validation (CV) with k = 5, repeated 10 times. The mean PA was calculated as the Pearson correlation between

observed and predicted trait values for the test data and recorded for each iteration. Pseudorandomization was used to divide the data into separate CV groups, with a fixed seed to ensure reproducibility. For each combination of marker set and model, we assessed the normality of the variance in the mean PA estimates from the 10 iterations, checked the distribution of the estimates, and then conducted an unpaired t-test (null hypothesis: the means of the two populations are equal).

Genomic relationship matrices (GRM)

In the tetraploid data, GRMs were constructed using *A.mat* function in the rrBLUP package (Endelman 2011). For the diploid data, GRMs and Principle Components (PCs) based on those GRMs were constructed using *calcG* function in KGD: Software for GBS-based relationship calculations v1.2.2 (Dodds et al. 2015).

Language editing

ChatGPT (GPT-4o, OpenAI's large-scale language model) was used for language editing. The edits were reviewed and revised by the authors, who take full responsibility for the final content of this publication.

Results

Comparison of marker sets and their performance on Prediction Accuracy for fry colour in a tetraploid potato panel

The first goal of this study was to evaluate different marker sets from multiplex amplicon sequencing data generated with PotatoMASH: individual SNPs, haplotypes identified through prior SNP calling (*haplotype-sites*), and haplotypes directly derived from read data (*haplotype-window*). We aimed to assess their performance in GP of 'of-the-field' (OTF) fry colour, using the previously published GBS-based prediction from Byrne et al. (2020) as a benchmark. In the original study (Byrne et al. 2020), GBS-SNPs with more than 10% missing data and clones with fewer than 10,000 data points were removed due to low sequencing coverage, resulting in a final dataset of 456 clones with 43.6k GBS SNPs. In the current study, all 607 clones had sufficient PotatoMASH data, reflecting the high sequencing depth and reliable locus recovery typical of amplicon-based genotyping, which yielded 2236 SNPs and 2000 haplotags (reported in Suppl. File 2). These datasets were used to compare the GP performance of the different PotatoMASH marker types against the GBS reference.

In addition, we utilised a newer haplotag caller within SMAP software (*haplotype-window*) to generate the third dataset for our study. *Haplotype-window* identified 3390 haplotags across the panel (reported in Suppl. File 2), ranging from 1 to 20 haplotags per locus in the population, with an average of 10 unique haplotags per locus, nearly doubling the average of 6 haplotags per locus identified using SMAP *haplotype-sites* on the same input data (Table 2, Fig. 2).

Using SMAP *haplotype-window*, we identified haplotags across 338 loci, which included 14 loci that had not previously shown any SNPs due to either limitations in variant calling or

stringent SNP filtering and were therefore classified as non-polymorphic by SMAP *haplotype-sites*. Only one of the 339 PotatoMASH loci remained non-polymorphic. Unlike SMAP *haplotype-sites*, which disregard pre-called indels sites and codifies haplotags at a given locus as same length based on the length covered by pre-called SNPs, *haplotype-window* uses the read sequence as the haplotype ID. Upon examining the read lengths, we found 154 loci had reads with varying lengths, indicating the presence of multiple indels variants per locus (ranging from 2 to 10 different read lengths per locus), while the remaining 184 loci had uniform lengths, suggesting no indels. Interestingly, within the 184 loci with uniform length, *haplotype-window* identified a higher number of haplotags than *haplotype-sites* in 164 of those loci. This suggests that *haplotype-window* also captures more genetic diversity unrelated to indels and is better at harnessing the full potential of the genetic diversity provided by PotatoMASH.

The minor allele frequency spectrum of the haplotags (MAF profile) of the four datasets used in this study reveals distinct differences between SNPs and haplotags (Fig. 3). In the SNP dataset (either generated by GBS or PotatoMASH), SNPs with MAF < 10% were filtered out due to concerns over potential sequencing errors, resulting in a more restricted distribution of alleles. In contrast to SNPs, which typically involve a major and minor allele at each locus, haplotypes represent multi-allelic variation, with each allele having its own frequency. As a result, haplotag datasets exhibited a broader and more complex allele frequency spectrum across the population. As per the haplotag sets, a notable shift in MAF profile was observed: for haplotags identified by SMAP *haplotype-window*, 1448 haplotags had a MAF below 1%, compared to 320 haplotags in the *haplotype-sites* dataset.

This pattern is likely driven both by true underlying allelic diversity and the sensitivity of the *haplotype-window* module to rare haplotypes. In particular, combinations of common SNP alleles can form rare multi-allelic haplotypes when constituent SNPs occur together infrequently. To minimize the inclusion of sequencing artifacts, only haplotypes supported by at least 10 sequencing reads and detected in a minimum of 10 individuals were retained (see Methods). Such patterns are not captured when filtering SNPs individually by MAF before haplotype construction. This difference in MAF profiles indicates that *haplotype-window* captures a broader range of genetic diversity, particularly regarding low-frequency alleles, compared to *haplotype-sites*. This enhanced diversity is attributed not only to the additional alleles captured with the *haplotype-window* module but also to the prior filtering of SNPs before haplotype calling with *haplotype-sites*.

Table 2. Summary differences of the output of the two haplotyping approaches with SMAP

Haplotags set	Number of haplotags	Missing data	Number of alleles per locus	Number of polymorphic loci
SNPs-based haplotags (with SMAP *haplotype-sites*)	2000	16%	2-14 (6 average)	325
Sequence-based haplotags (with SMAP *haplotype-window*)	3390	24%	1-20 (10 average)	338

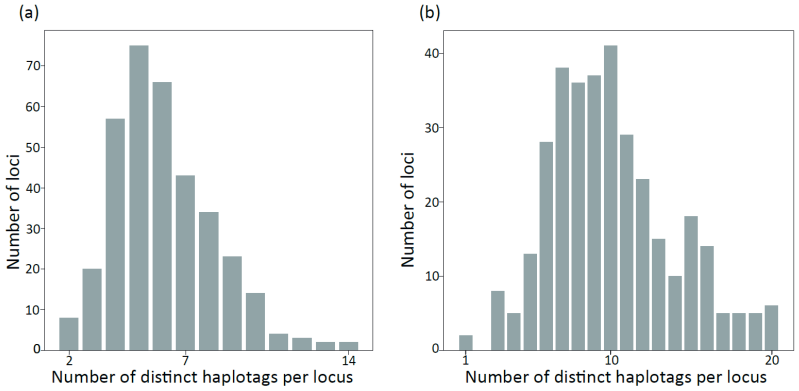


Fig. 2 *Distribution of the number of distinct haplotags per locus across PotatoMASH loci in the tetraploid panel (a) Distribution of the 2000 haplotags from SMAP haplotype-sites in 325 loci (b) Distribution of the 3390 haplotags from SMAP haplotype-window in 338 loci.*

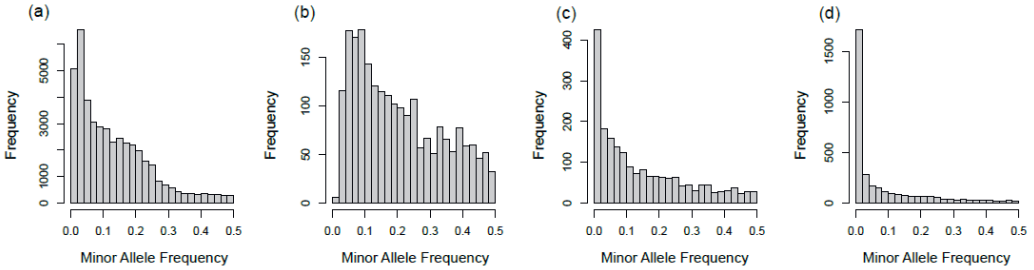


Fig. 3 *Minor Allele Frequency (MAF) distribution for a) 43.6k GBS SNPs b) 2236 PotatoMASH SNPs c) 2000 haplotags generated with haplotype-sites d) 3390 haplotags generated with haplotype-window.*

To assess the ability of the four datasets to represent genetic relationships among individuals, GRMs were constructed for the tetraploid potato panel. The heatmaps derived from these GRMs showed similar patterns of relatedness, with individuals from the three breeding years exhibiting mixed relatedness to one another (Fig. 4). Despite using different

marker sets (GBS and PotatoMASH data, specifically), the genetic relationships among the individuals remain consistent, with no clear distinction in relatedness patterns between clones from the three years of the breeding programme.

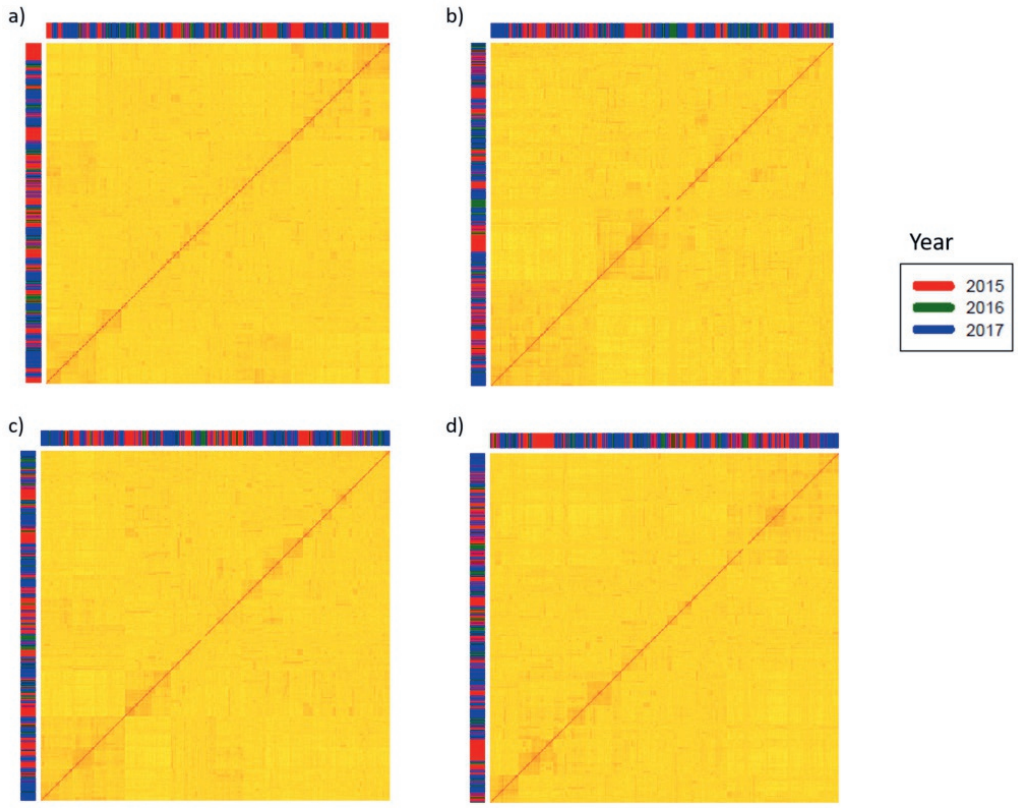


Fig. 4 *Genomic relationship matrices (GRM) heatmaps of the 456 tetraploids potato clones (the subset used in Byrne et al. 2020) based on four marker sets: a) GBS SNPs, b) PotatoMASH SNPs, c) haplotags from SMAP haplotype-sites, and d) haplotags from SMAP haplotype-window. The heatmap colour key indicates the strength of genetic relatedness, with red representing high genetic similarity and yellow indicating lower genetic relatedness. The lateral colour palette represents the breeding material from three years of the breeding programmes, highlighting the distribution of clones from each year within the genetic relationship matrix.*

To quantify the similarity of genomic relationship estimates between datasets, particularly comparing the estimates obtained from GBS data to those from PotatoMASH, we performed a linear regression analysis between variables from pairwise matrices. Fig. 5 shows a matrix plot with all pairwise comparisons of the four GRM estimates, focusing on the relatedness values (off-diagonal elements). The lower diagonals display scatter plots of GRM estimates between each pair of marker sets, while the upper diagonals show the corresponding regression results. Although the correlation matrix analysis (Fig. 5) reveals a small bias in genetic relationship estimates between GBS-derived SNPs and PotatoMASH markers, the four pairwise comparisons of genetic relationship matrices show high correlations between the genomic relationships estimated from GBS data and those from the PotatoMASH

datasets (SNPs and haplotags), with correlation coefficients ranging from 0.84 to 0.88. This demonstrates that the 338 loci from PotatoMASH capture genomic relationship estimates with remarkable accuracy, performing almost as well as the 43.6k GBS-based SNPs despite being a fraction (about 5%) of the markers. While PotatoMASH targets 339 fixed loci, GBS typically samples hundreds of thousands to over a million loci across the genome, depending on sequencing depth and filtering (Byrne et al. 2013; Elshire et al. 2011).

Additionally, comparisons within the PotatoMASH datasets themselves showed high regression coefficients, with the highest correlation between the *haplotype-window*-based GRM and the *haplotype-site*-based GRM estimates (r = 0.956), suggesting almost no difference in performance between the two haplotype sets for genomic relationship estimation.

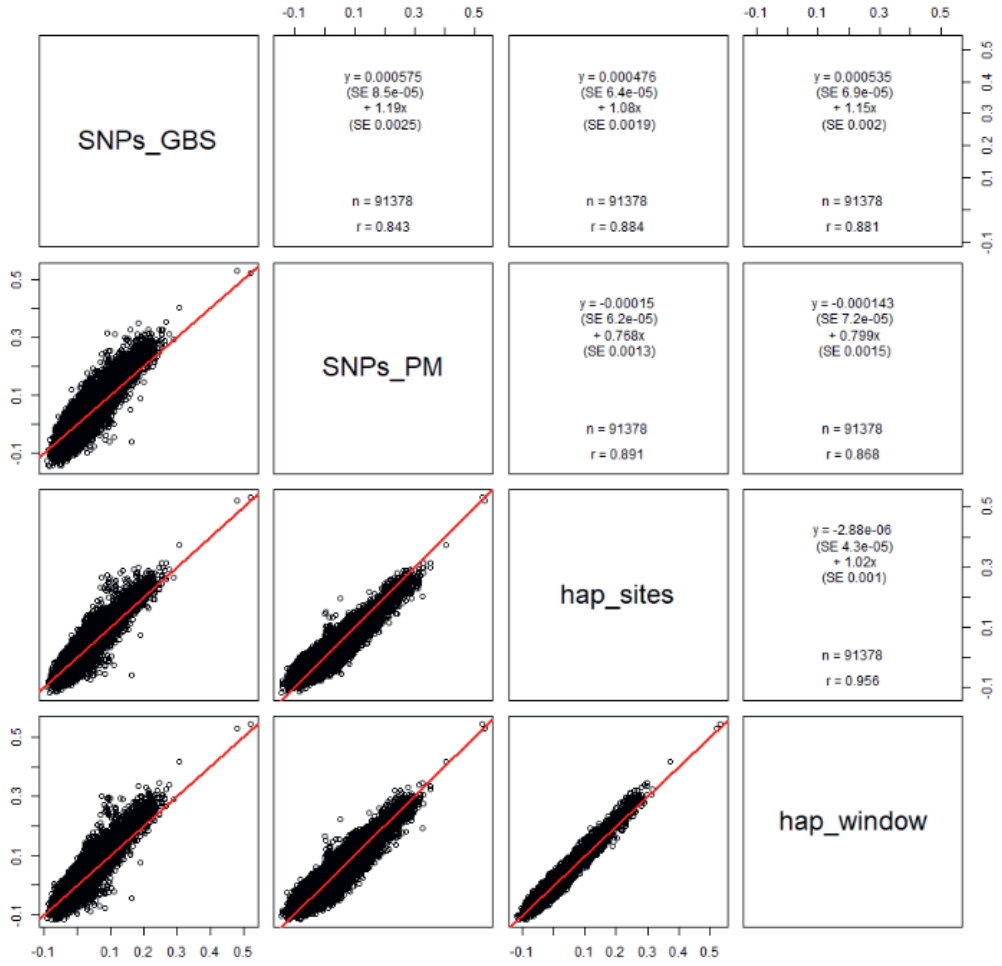


Fig. 5. *Matrix plot comparing the relatedness estimates (off-diagonal) between the different Genomic relationship matrices (GRM). The lower diagonals display scatter plots of relatedness estimates for each pair of GRMs, and the upper diagonals show the corresponding regression output. A red line is drawn to indicate*

where the values are equal, providing a visual reference for the strength of agreement between relatedness estimates across the four marker sets.

To compare the PA of different approaches for analysing PotatoMASH data to the PA achieved with GBS high-density marker set, we applied the same GP framework and the same GP models used by Byrne et al. (2020). In that study, the GP models were first developed using data from each year and then applied to predict OTF fry colour in the remaining years, rotating each year as the training population.

The highest PA achieved using PotatoMASH was 0.70, compared to 0.77 for the 43.6k GBS SNP set. The mean PA achieved with PotatoMASH data ranged from 0.20 to 0.69 for SNPs and 0.29 to 0.70 for either haplotag set (Table 3). When comparing these results with the PA reported by Byrne et al. (2020) for the 43.6k GBS SNP set (PA ranging from 0.11 to 0.77), we observed, on average, a 14% decrease in PA for the PotatoMASH SNPs and a 9% decrease for either haplotag set when using the rrBLUP, BayesA, and BayesLASSO algorithms. However, for the Random forest algorithm, the PA with haplotags was either the same or higher (reported in Suppl. File 3). Consistent with the results of Byrne et al. (2020), our findings showed minimal variation in PA across the different GP algorithms, with the exception of Random Forest, which consistently showed lower PA. Lower PA was also observed when using the 2016 data as the training or test population, which was previously attributed by Byrne et al. to reduced relatedness between the material from that year and other years. Although PotatoMASH data showed overall lower PA compared to the 43.6k SNP set, in most cases, this decrease was minimized when using haplotags from *haplotype-window*, which almost consistently achieved higher PA than the SNP-based haplotags from *haplotype-sites* (Table 3). This suggests that *haplotype-window* offers a more effective approach in maximizing the PA of PotatoMASH data. Compared to GBS data, a larger bias was observed when using the 2016 data as the training or test population with PotatoMASH data. This bias is likely due to the combination of reduced relatedness between the training and test populations, along with the low-density markers in the PotatoMASH dataset.

Table 3. The prediction accuracy (PA) of all marker sets for 'off-the-field' fry colour in the tetraploid FRY population, using various combinations of training and test populations with rrBLUP (bias is shown in brackets). Detailed breakdown of the results for all statistical algorithms used is reported in Suppl. File 3.

Training Set	Test Set	GBS SNPs (43.6k total)	PotatoMASH SNPs (2236 total)	haplotype-site (2000 total)	haplotype-window (3390 total)
2015	2016	0.26 (0.43)	0.33 (0.69)	0.32 (0.59)	0.32 (0.68)
2015	2017	0.75 (1.05)	0.67 (0.99)	0.67 (1)	0.67 (0.94)
2017	2015	0.77 (1.29)	0.68 (1.01)	0.68 (1.08)	0.7 (1.14)
2017	2016	0.48 (1.05)	0.31 (0.54)	0.38 (0.75)	0.38 (0.73)
2016	2017	0.56 (3.26)	0.42 (7.3)	0.46 (6.71)	0.48 (6.99)
2016	2015	0.49 (2.59)	0.45 (7.93)	0.48 (6.99)	0.41 (5.79)

Genomic Prediction for multiple traits in diploid potato

A second objective of this study was to evaluate the use of haplotags identified from multiplex amplicon sequencing in GP and determine any added value of using short haplotypes over SNP data in GP of traits routinely evaluated in potato breeding. To do this, we used a panel of 558 diploids genotyped with PotatoMASH and evaluated for 23 agronomic and quality traits (Vexler et al. 2024). In order to capture the maximum genetic information for GP, we used less stringent filtering settings compared to the previous genotyping and included all SNPs with a MAF higher than 1%, resulting in a dataset of 2730 SNPs and 2955 haplotags based on those SNPs called with SMAP *haplotype-sites* (Vexler et al. 2024). In this study, haplotags were also called using SMAP *haplotype-window*, identifying a total of 5919 haplotags across the panel (reported in Suppl. File 4), ranging from 3 to 38 haplotags per locus, with an average of 17 unique haplotags per locus. This nearly doubles the average of 9 haplotags per locus identified by SMAP *haplotype-sites* on the same input data (Table 4, Fig. 6). For *haplotype-window*, 201 loci had haplotypes of varying lengths, indicating multiple indel occurrences, ranging from 2 to 18 read lengths per locus, while 138 loci had uniform lengths, suggesting no indels. Like what was observed in the tetraploid FRY population, for all 138 loci with single length, a higher number of haplotags per locus were observed when haplotags were called with *haplotype-window*, indicating that *haplotype-window* captures more genetic diversity unrelated to indels compared to *haplotype-sites*. Additionally, the MAF profiles across all loci showed distinct shifts for the three datasets (Fig. 7). After filtering, only 43 SNPs had a MAF below 1% while 928 haplotags from SMAP *haplotype-sites* and 3419 from *haplotype-window* had MAFs below 1%.

Table 4. Summary differences of the final output of the two haplotyping approaches with SMAP

Haplotags set	Number of haplotags	Missing data	Number of haplotags per locus	Number of polymorphic loci
SNPs-based haplotags (with SMAP *haplotype-sites*)	2955	7%	2-30 (9 average)	334
Sequence-based haplotags (with SMAP *haplotype-window*)	5919	9%	3-38 (17 average)	339

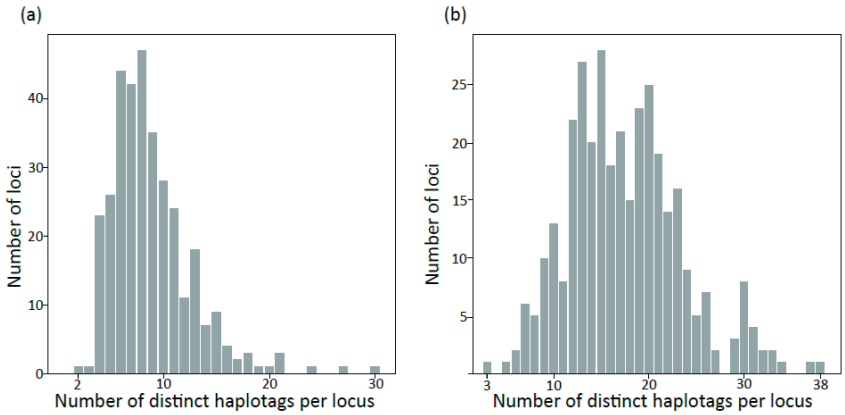


Fig. 6. *Distribution of the number of distinct haplotags per locus across PotatoMASH loci in the diploids panel (a) for the 2955 SNP-based haplotags from SMAP haplotype-sites (b) for the 5919 sequence-based haplotags from SMAP haplotype-window.*

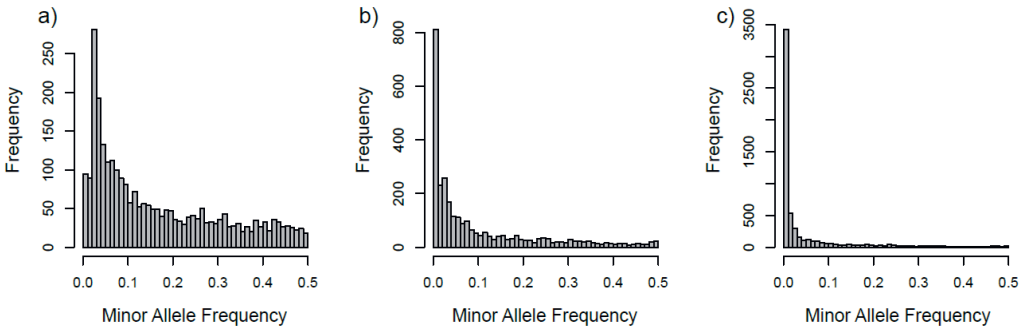


Fig. 7*. Minor Allele Frequency (MAF) distribution for a) 2730 SNPs, b) 2995 SNPs-based haplotags from SMAP haplotype-sites, c) 5919 sequence-based haplotags from SMAP haplotype-window.*

GRMs were constructed for the diploid potato panel using the three distinct marker sets. For calculating these GRM, markers not shared between at least two individuals were excluded from the analysis, resulting in 2224 haplotags called from *haplotype-sites* and 2635 haplotags from *haplotype-window*, while none of the SNPs were removed as they were all shared with at least two individuals. The GRMs were visualized through heatmaps (Fig. 8). Notably, individuals from the Meijer breeding programme (represented in pink) were genetically distinct from individuals from other breeding programmes, forming a separate cluster. In the SNP-based GRM, the individuals from the Meijer programme split into two sub-clusters.

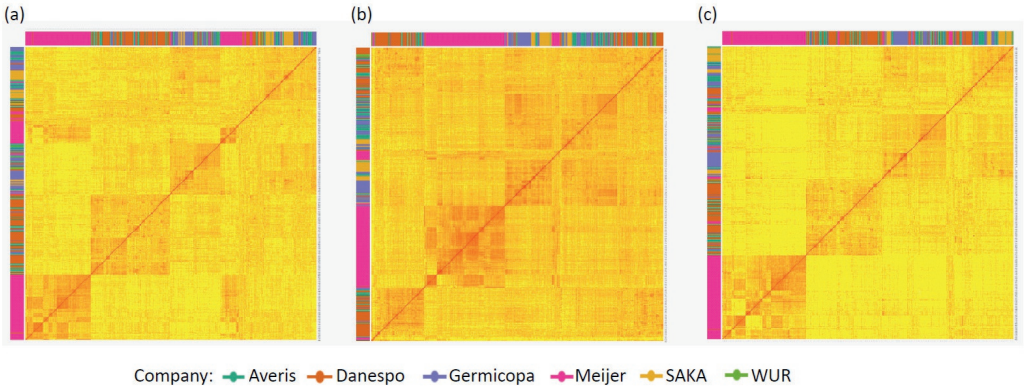


Fig. 8 *Genomic Relationship Matrices (GRM) heatmap with hierarchical clustering of the 558 diploid potato clones based on three marker sets: a) SNPs, b) haplotags from SMAP haplotype-sites, and c) haplotags from SMAP haplotype-window. The heatmap colour key indicates the strength of genetic relatedness, with red representing high genetic relatedness and yellow indicating lower genetic relatedness. The lateral colour palette represents the six breeding programmes, highlighting the distribution of clones from each programme within the GRM.*

Additionally, Principal Component Analysis (PCA) was performed of the GRMs based on the first two principal components (Fig. 9). The PCA plots demonstrated two main clusters, with the Meijer sub-population deviating from the other breeding programmes. The variance explained by the first principal component (PC1) is 31.15% for the SNPs, 33.49% for the SNP-based haplotags from *haplotype-sites*, and 35.13% for the sequenced-based haplotags from *haplotype-window* and Principal Component 2 (PC2) explains 15.15%, 15.56%, and 14.02% of the variance for each respective marker set. Both PCs explain a substantial portion of the total variation in the data. The slight variations in the variance explained by the principal components across the marker sets still showed similar population structures, emphasizing the utility of PotatoMASH in capturing the genetic relationships and diversity within a multi-programme diploid potato breeding panel.

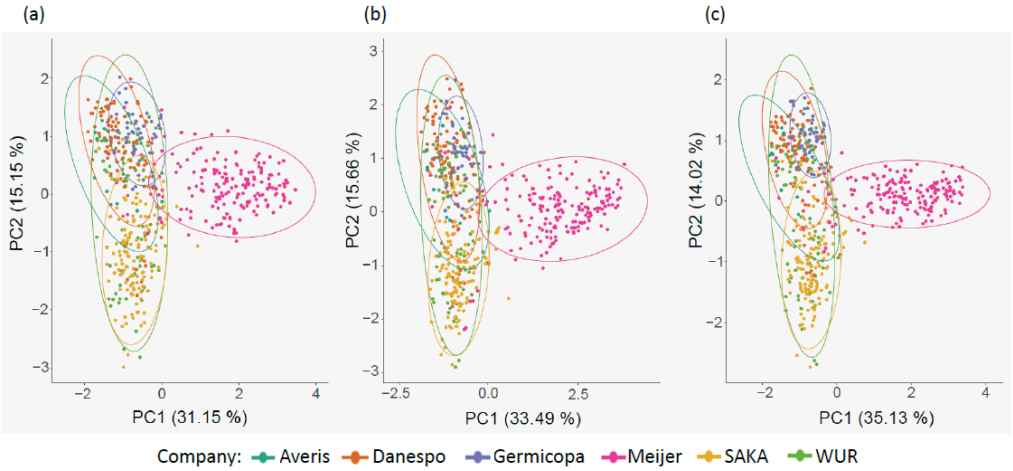


Fig. 9 *Population structure with Principal Component Analysis (PCA) of the 558 diploid potato clones based on the Genomic Relationship Matrices (GRM) derived from the three PotatoMASH marker sets: a) SNPs, b) haplotags from SMAP haplotype-sites, and c) haplotags from SMAP haplotype-window. The first two Principal Components (PC1 and PC2), shown on the axes, with with the percentage of variance they explain indicated in parentheses. Colours represent breeding companies, and ellipses indicate 95% confidence intervals for each group.*

This panel was phenotyped for 23 agronomic and quality traits, and control varieties were used to estimate the BLUEs of phenotypic means across years and locations (Vexler et al., 2024). We used the phenotypic data to evaluate the PA for these traits using our three marker sets. We observed moderate to high mean PA values (0.29–0.81) for all 23 traits, with small variations in PA across the three marker sets and between the models tested (Fig. 10). In Fig. 10 the scales of individual graphs were set to highlight differences. To standardise the scale, the results are also presented with a fixed y-axis (ranging from 0 to 1) in Suppl. File 6.

Overall, differences in PA between SNPs and haplotags were small for most traits and models, though many of these small differences in mean PA were still statistically significant.

Haplotags outperformed SNPs for 11 traits, including Canopy Stage 1, Canopy Stage 2, Tuber Length, Total Tuber Number, Tuber Regularity, Yellow Skin Colour, Presentability of Tubers, Sensitivity to Common Scab, Enzymatic Browning, Dry Matter Content, and Sprout Dormancy. SNPs performed better for six traits: Yield, Skin Smoothness, Cooking Type, After-Cooking Blackening, Chipping Colour 1_8, and Chipping Colour 2_8. For Tuber Shape, Yellow Flesh Colour, Eye Depth, Skin Brightness, Chipping Colour 2_4, and Maturity, PA results were mixed across models or showed no clear advantage for either marker type.

Among the 11 traits where higher PA was obtained with the haplotags, six performed best with haplotags from SMAP *haplotype-window*: Canopy Stage 1, Total Tuber Number, Tuber Regularity, Yellow Skin Colour, Presentability of Tubers, and Dry Matter Content. Two traits,

Tuber Length and Sprout Dormancy, had higher PA with haplotags from SMAP *haplotype-sites*. The remaining three traits showed similar PA between both haplotype-based methods. Full PA results for all models are provided in Suppl. File 5.

The lowest mean PA was observed from the rrBLUP mode for Canopy Stage 2 and Total Tuber Number (0.29 and 0.33 respectively), while the highest PA was observed with the Bayes B model for Tuber Length (0.79) and Yellow Flesh Colour (0.76). BayesB model and BayesC generally performed better than rrBLUP, GBLUP, and BRR models, particularly in Canopy Stage 2, Tuber Shape, Yellow Flesh Colour, Eye Depth, Chipping Colour 2_8, and Maturity. When transitioning from SNPs to SNPs-based haplotags from SMAP *haplotype-sites*, an increase in PA was observed for the majority of traits. However, Yield, Cooking Type, After-Cooking Blackening, and Chipping Colour 1_8 showed better PA with SNPs.

Significant increases in PA were observed for certain traits with the haplotags called by SMAP *haplotype-sites* compared to individual SNPs that composed those haplotags, for example: Yellow Skin Colour (4% increase), Dormancy (5% increase), Sensitivity to Scab (6% increase), and Dry Matter Content (13% increase). The sequence-based haplotags from SMAP *haplotype-window* provided even larger increases in PA compared to the SNPs, such as 8% for Yellow Skin Colour, 9% for Enzymatic Browning, and 15% for Dry Matter Content. However, there were some cases where the transition from SNPs to haplotags resulted in a decrease in PA, such as a 7% decrease for Yellow Skin Colour and a 9% decrease for Yield.

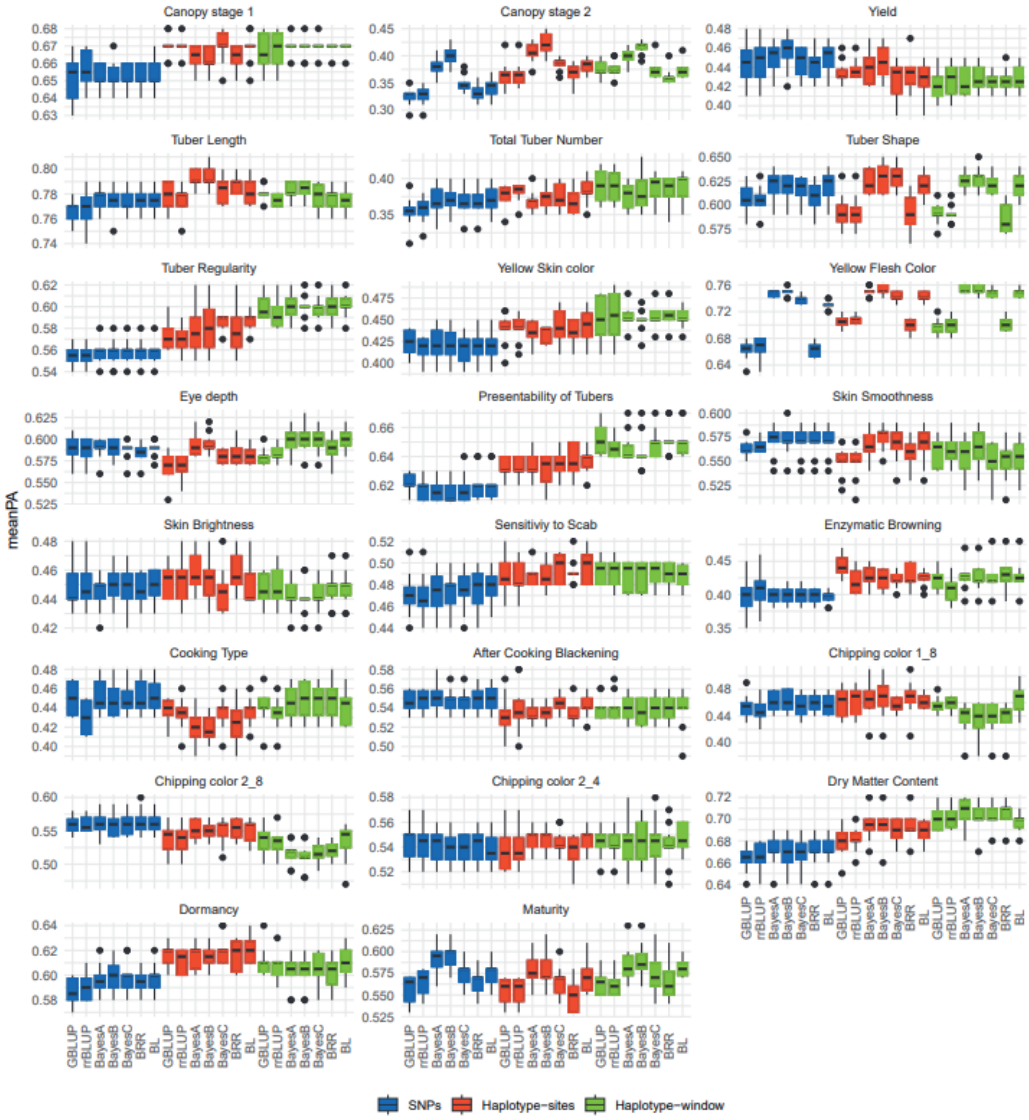


Fig. 10 *Effect of different marker sets on the mean prediction accuracy (PA) for 23 agronomic and quality traits in diploid potato. For each trait the mean PA obtained is plotted from left to right for seven different models using SNP data (in blue, on the left), haplotags identified by haplotype-sites (in red, in the middle) and haplotags identified by haplotype-window (in green, on the right). The Mean PA is calculated as the average Pearson correlation obtained in a k = 5-fold cross-validation scenario, repeated over 10 iterations. Detailed breakdown of the results for all traits is reported in Suppl. File 5.*

A specific example for rrBLUP of the mean PA is presented in Table 5, which also includes the average estimated trait heritability for each breeding programme previously published in Vexler et al. (2024). The PA was consistently lower than the heritability for all traits, except for Canopy Stage 1, which low value of heritability was due to a particular breeding

programme (0.22), while material from the other programmes ranged from 0.54 to 0.85 (data not shown).

Table 5. Mean PA with rrBLUP for the 23 traits with the three marker sets and the broad sense (H^2) Heritability.

Trait	mean PA SNPs	mean PA *haplotype-sites*	mean PA *haplotype-window*	Average H^2 across all companies (Vexler et al. 2024)
Canopy Stage 1	0.66	0.67	0.67	0.60
Canopy Stage 2	0.33	0.36	0.38	0.75
Yield	0.45	0.44	0.43	0.86
Tuber Length	0.77	0.78	0.78	0.87
Total Tuber Number	0.36	0.38	0.39	0.69
Tuber Shape	0.61	0.59	0.59	0.90
Tuber Regularity	0.55	0.57	0.59	0.65
Yellow Skin Colour	0.42	0.44	0.45	0.64
Yellow Flesh Colour	0.67	0.71	0.7	0.89
Eye depth	0.59	0.57	0.58	0.83
Presentability of Tubers	0.62	0.63	0.65	0.75
Skin Smoothness	0.57	0.55	0.56	0.60
Skin Brightness	0.45	0.45	0.45	0.62
Sensitivity to Common Scab	0.47	0.49	0.49	0.51
Enzymatic Browning	0.41	0.42	0.41	0.82
Cooking Type	0.43	0.43	0.43	0.74
After-cooking Blackening	0.55	0.54	0.54	0.73
Chipping Colour 1_8	0.45	0.46	0.46	0.78
Chipping Colour 2_8	0.56	0.54	0.53	0.84
Chipping Colour 2_4	0.54	0.54	0.54	0.82
Dry Matter Content	0.66	0.68	0.7	0.89
Sprout Dormancy	0.59	0.61	0.61	0.75
Maturity	0.56	0.56	0.56	0.82

Discussion

Marker Assisted Selection and Genomic Selection are powerful approaches in plant breeding, utilizing genetic data to enhance the selection of superior offspring. However, as noted in the introduction, methods such as GBS for genome-wide genotyping can be costly, and SNP assays are often affected by ascertainment bias due to the need for a development phase (Vos et al. 2015). Alternative marker systems, such as KASP, have been shown to be cost effective for applications such as MAS, but are impractical to the selection of complex traits.

In response, we propose an enhanced breeding strategy employing a cost-effective, genome-wide marker system based on pooled multiplex amplicon sequencing (referred to as PotatoMASH), to facilitate genomic-assisted breeding. In this study, we demonstrate that, despite generating fewer markers, PotatoMASH efficiently captures genetic diversity and provides prediction accuracy broadly comparable to GBS approaches based on reducing genome complexity with restriction enzymes. Moreover, we validate its efficacy across a wide range of traits commonly selected in potato breeding, underscoring its versatility and potential to enhance breeding efficiency.

PotatoMASH: An integrated marker system for MAS and GS in the breeding cycle

As outlined in the introduction, numerous studies have demonstrated the feasibility of GP in potato breeding, but few have considered its practical deployment from the point of view of affordability and how it fits into the logistical framework of an ongoing commercial breeding programme.

A typical potato breeding process, based on the Teagasc-IPM Potato Breeding Programme, is illustrated in Fig. 11a. It is based on the widely used "early generation intensive selection" model, which operates within a structured 12-year cycle of trials and selections. The process begins with initial crosses in Year 1, followed by heavy selection pressure at the first field stage, where poor-performing genotypes are quickly eliminated. This allows for more intensive selection for a smaller subset of individuals. As the programme progresses, selection pressure drops as the number of variety candidates is reduced, but the scale and complexity of evaluation (replication, environmental testing, and phenotyping) increase.

A key component of the breeding scheme outlined in Fig. 11a is the use of KASP markers to screen and select for resistant material. As new resistance targets are identified, they are converted into KASP markers and incorporated into the selection panel for MAS. Currently, in the Teagasc-IPM programme, MAS is applied in Year 4 to identify clones carrying disease-resistant alleles, at which point approximately 2500 clones remain in the selection pipeline. Individuals containing resistant loci (R-loci) can continue through the programme for potential variety selection. However, the most promising genotypes can also be reintroduced as parents at any stage, accelerating recurrent selection and allowing the accumulation of multiple R-genes. This strategy reduces the time between breeding cycles,

increasing the efficiency of resistance breeding. While this approach has been highly successful in producing high-performing resistant varieties, it does not improve selection for complex traits.

We propose an enhanced breeding strategy incorporating PotatoMASH into population improvement and product development for routine cost-effective genotyping, thereby facilitating combined MAS and GS on a single platform (Fig. 11b). The initial phase comprises a 4-year population improvement cycle. This encompasses initial crosses and a first field generation that typically involves the elimination of poor performing plants resulting from the high genetic load. In the second field generation (at year 4), a single round of genotyping using PotatoMASH facilitates both MAS and GS to identify superior lines, with GS enabled by the genome-wide coverage and MAS enabled by the addition of specific targeted markers, for instance at the resistance loci formerly screened by individual KASP markers. Selected high genetic merit lines are subsequently utilised for the next cycle of crosses in the population improvement phase, as well as being advanced for subsequent product development, where those lines are evaluated in multi-location field trials. Ultimately, the evaluation data from the population improvement cycle will inform the GP models with additional phenotypic data to increase the prediction accuracy over successive cycles (Fig. 11b). Over time, as sufficiently large training data sets build up, more traits will become accessible to GS.

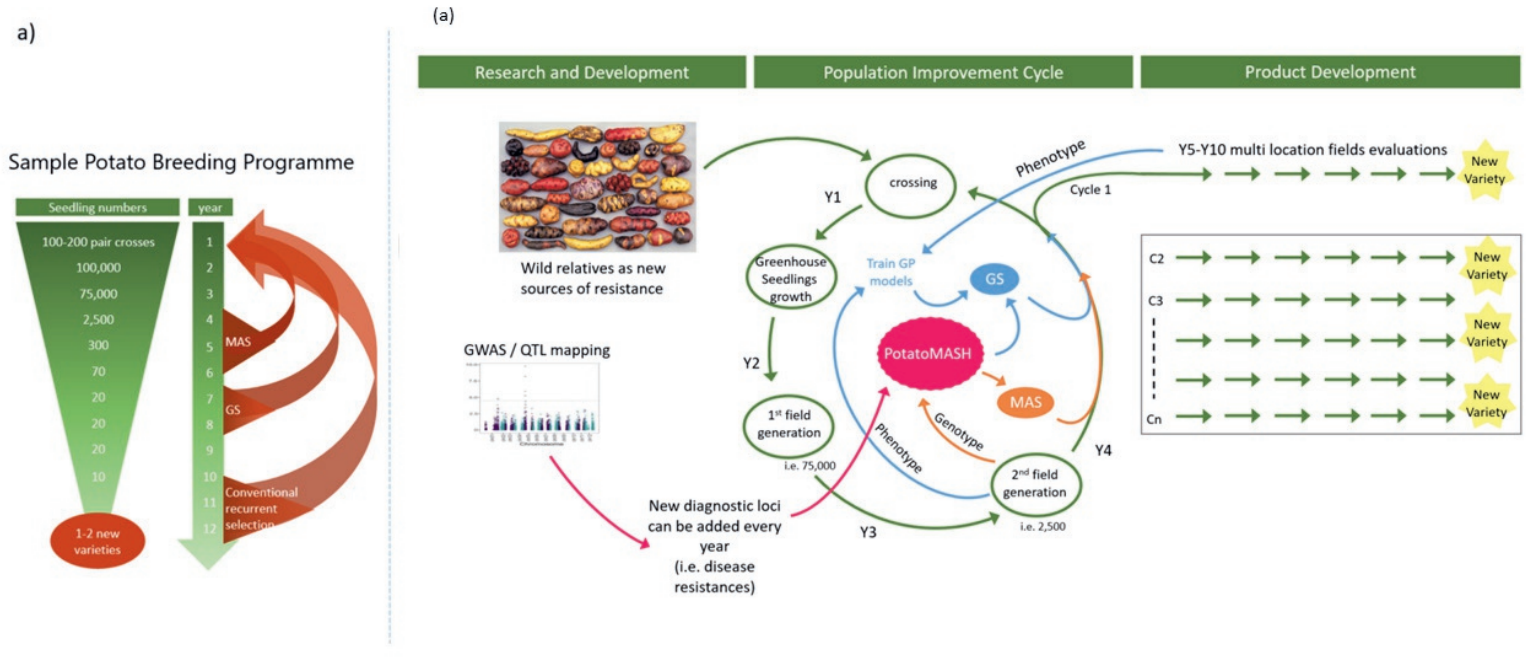


Fig. 11 *a) A sample potato-breeding programme typically for tetraploid potato breeding. b) An enhanced breeding strategy using PotatoMASH, structured into three stages. The central phase consists of a 4-year population improvement cycle. In the second field generation, a single round of genotyping with PotatoMASH enables both Marker-Assisted Selection (MAS) and Genomic Selection (GS) to identify best-performing lines. These lines contribute to the next breeding cycle and advance to product development through multi-location field trials. Ongoing research in the first phase allows the continuous integration of new loci into the marker system.*

A research and development (pre-breeding) phase, illustrated on the left side of the enhanced breeding strategy in Fig. 11b, is a common and essential component of most breeding schemes. This phase typically involves marker-trait association studies, driven by QTL detection, to inform downstream selection strategies. However, as outlined in the Introduction, many agronomic traits in potato are genetically complex and environmentally sensitive. These traits often involve subtle, polygenic signals that are difficult to model mechanistically or resolve through traditional QTL mapping. In addition, phenotyping such traits for selection is labour-intensive, time-consuming, and costly, especially under multi-environment trials or stress conditions. In this context, genomic prediction acts as a generalised “black-box” solution, enabling selection based on genome-wide patterns rather than prior knowledge of trait architecture. The prediction success of PotatoMASH for complex traits, not only quantitatively measured traits like yield and chipping colour, but also visually assessed traits such as Tuber Regularity and Presentability of Tubers (PA = 0.50–0.66, Fig. 10), demonstrate the value of genomic selection as a scalable strategy for traits that are difficult to phenotype reliably or are prone to subjective bias.

One important aspect of this strategy is the ease with which the genotyping assay can be updated over time. As novel germplasm enters the breeding programme, it is likely to be useful to add markers to capture novel single locus targets and to improve marker density. As demonstrated by Leyva-Pérez et al. (2022), PotatoMASH allows for the addition of diagnostic markers linked to traits to the core set of 339 loci with relative ease, allowing “evolution” of the platform over time.

PotatoMASH performance in Genomic Prediction

Given the application goals described above, the first objective of this study was to test the potential of a low-density genotyping tool like PotatoMASH for use in GP in comparison to a more standard high-density genotyping platform of the type that has been widely used to date. We compared differences in prediction accuracy for the moderately high heritability trait OTF fry colour with PotatoMASH against predictions made with high-density GBS data generated for the same tetraploid population (Byrne et al. 2020). In this population, Leyva-Pérez et al. (2022) found that PotatoMASH haplotags could identify a QTL for fry colour, whereas the SNP set from which they were derived could not. Therefore, for PotatoMASH predictions, we examined both SNP and haplotag derived predictions.

The PA results for OTF fry colour (Table 3) show that PotatoMASH yields promising PA for GS, despite using far fewer markers than GBS. PotatoMASH generates 2236 SNPs, 2000 SNP-based haplotags, and 3390 sequence-based haplotags, substantially fewer than the 43.6k SNPs from GBS. Despite this reduction, the average decline in PA is only 14% with PotatoMASH SNP data and 9% with haplotag data, compared to GBS SNPs. Given that PotatoMASH genotyping costs 10 to 20 times less per clone than GBS, this modest PA reduction is a favourable trade-off. This suggests that PotatoMASH could constitute a cost-

effective alternative for breeding programmes needing to genotype thousands of clones annually for GS without significantly compromising PA, at least for high heritability traits.

Prediction Accuracy using SNP-based haplotags compared to the individual SNPs

At the outset of this study, we hypothesized that haplotags, by more accurately capturing the true allelic configuration of loci than individual SNPs, might lead to improved PA. In the tetraploid FRY population, we observed a general trend of haplotags yielding slightly higher PA than their corresponding SNPs across most models tested for a single trait, OTF fry colour. Building on this, we applied PotatoMASH for GP across 23 agronomic and quality traits in a diploid breeding panel, where PA values ranged from medium to high (0.29 - 0.81). While haplotags generally improved PA for certain traits, this advantage was not universal. In six traits, SNPs slightly outperformed haplotags, and overall, marker types exhibited minimal differences in PA (Table 5).

As mentioned in the introduction, this panel of 558 diploid clones was previously utilised for GWAS (Vexler et al. 2024), revealing differences in QTL detection performance between SNPs and haplotags. In the current study, we observed no association between PA and the ability of SNPs and haplotags to detect QTL for a trait. For instance, in (Vexler et al. 2024), no QTL were detected for Tuber Regularity, Skin Brightness, and Presentability of Tubers. However, in the present study, moderate to high PA were observed for these traits, ranging from 0.50 to 0.66 (Fig. 10). For Tuber Shape, despite a highly significant QTL on chromosome 10 (-log10(P-value) = 14.36), the PA was comparable to polygenic traits, ranging from 0.57 to 0.65. There was also no consistent association between QTL discovery performance of a particular marker type (SNPs or haplotags) and its performance for GP in this study. For example, in After-Cooking Blackening, a QTL was detected with SNPs on chromosome 3, and a higher PA was observed with SNPs. For Cooking Type, where no QTL was detected with SNPs but one was identified with haplotags on chromosome 8, the PA was still higher with SNPs. For Dry Matter Content, where a QTL was identified with SNPs, the PA was higher with haplotags. For Skin Smoothness, where a QTL was detected with haplotags on chromosome 6, the PA was higher with SNPs. These findings suggest that the relationship between QTL detection and PA is influenced by trait-specific genetic characteristics, marker type, and the effect size of markers in a given population.

Haplotype-window captures more genomic diversity

Subsequent to the GWAS study undertaken by Vexler et al (2024), the SMAP software adopted for read-backed haplotyping was updated. The original SMAP software employed an approach called *haplotype-sites*, which reconstructs haplotags based on the polymorphisms it detects at pre-called SNP positions (sites) in mapped reads. An additional module, *haplotype-window* was originally designed to call indels for molecular biology applications, such as identifying induced mutations from CRISPR (Develtere et al. 2023; Lorenzo et al. 2023). Similar to the *k-mer* concept, *haplotype-window* treats sequence signatures as distinct entities, which we can interpret as haplotags. In the rationale

underlying SMAP *haplotype-window*, the entire DNA sequence between two fixed 'borders' are extracted per read and each unique haplotype detected is considered a haplotag. For practical reasons, the eight 3'-nucleotides of the forward and reverse primer per amplicon are taken as 'border' sequences that define the locus. Therefore, in addition to SNP-based variation, *haplotype-window* can also capture combined SNP and indel variation, without the dependency of *a priori* SNP and indel calling with software such as GATK or SAMtools, and subsequent SNP filtering. We investigated whether this approach would significantly increase haplotag number per locus and capture a broader spectrum of allele diversity, including genetic variants not represented in SNP-based haplotag sets, and whether the increased information content could enhance GP accuracy. In the tetraploid FRY population, sequence-based haplotags from SMAP *haplotype-window* generated a larger set (3390) of haplotags compared to SNP-based haplotags from SMAP *haplotype-sites* (2000). The average number of unique haplotags per locus increased from 6 with SNP-based haplotags to 10 with sequence-based haplotags from *haplotype-window* (Table 2, Fig. 2). At least some of this increase resulted from indels, as evidenced in the observation of length polymorphism at ~45% (154/339) of loci. However, *haplotype-sites*, unlike *haplotype-window*, depends on the set of SNPs used as input. In this population, to maintain compatibility to what was previously done by Leyva-Pérez et al. (2022) we filtered SNPs prior to haplotyping using a stringent set of parameters. Because of this, many SNPs were excluded, resulting in fewer haplotags from *haplotype-sites*. Theoretically, with PotatoMASH's high sequencing depth, SNP filtering may not be necessary; the mapped read data can be directly used to call haplotypes with SMAP *haplotype-sites*, with filtering applied only to haplotyping parameters. For the diploid panel, less stringent filtering settings were applied for SNP variant calling compared to the procedure used by Leyva-Pérez et al. (2022). While this approach risks false positives due to sequencing errors, using SNPs with high-quality sequencing data (average sequencing coverage of 813 reads per locus per sample; Vexler et al. 2024) provided confidence that capturing increased genomic diversity outweighed the risk. This resulted in 2730 SNPs yielding 2995 haplotags using *haplotype-sites* compared to 5919 haplotags when using *haplotype-window*. The average number of unique haplotags per locus increased from 9 with SNP-based haplotags to 17 with *haplotype-window* (Table 4 and Fig. 6). Interestingly, in this case, the proportion of loci exhibiting length polymorphisms increased to ~60% (201/339), probably as a reflection of a more diverse genetic origins of the diploid panel.

Our hypothesis proposed that sequence-based haplotags would enhance PA by incorporating additional genetic variants, such as indels, present in 45% of loci in the tetraploid population and 60% of loci in the diploid panel. These variants may be in LD with traits not captured by SNP-based markers. We anticipated that increasing the number of markers would augment the statistical power of the prediction models.

The results from the OTF fry colour prediction in the tetraploid population partly corroborated this hypothesis, as the PA achieved with sequence-based haplotags was, in most instances,

superior or equivalent to that obtained with SNP-based haplotags (Table 3). In the diploid panel, comparing PA results of the two haplotag sets across all 23 traits, 10 traits exhibited identical PA results. For the remaining traits, although differences were minor, higher PA was observed with haplotags derived from *haplotype-window* for 8 traits, while haplotags from *haplotype-sites* produced higher PA for 6 traits (Fig. 9).

A significant technical advantage of the *haplotype-window* approach over *haplotype-sites* is that it provides a stable identifier based on the actual sequence, independent of pre-called and filtered SNP sites, which depend on the population genotyped. This renders the sequence-based haplotag's identifier comparable between populations, enhancing dataset interoperability for various research and breeding applications. In contrast, the SMAP *haplotype-sites* software uses a string code that may vary across populations with different SNP sets, complicating the tracking of haplotags across populations.

Influence of genetic structure of trait – MAS or GS for low complexity traits?

For the diploid panel we tested seven different models in GP. For most traits, the differences in mean PA across the models were minimal. However, for traits that are primarily controlled by one or a few major genes, Bayesian methods, specifically BayesA, BayesB and BayesC, consistently outperformed GBLUP and ridge regression methods, providing higher PA. This was particularly evident for traits such as Tuber Shape/Eye Depth, Maturity/Canopy Stage 2 and Yellow Flesh Colour controlled mostly by single loci in chromosomes 10 (*Ro* locus) 5 (*StCDF* locus) and 3 (*Bch* locus) respectively (Vexler et al. 2024). The distinction between the models arises from their different approaches to marker effects. BLUP-based methods, which are linear models, assume that marker effects follow a normal distribution and that all markers contribute equally to the trait. In these models, the trait is assumed to be influenced by many QTL, each with a small effect on the trait. In contrast, Bayesian methods, particularly BayesA, BayesB and BayesC, are linear parametric models that incorporate mixture priors. These methods assume that all markers have an effect, but the magnitude of these effects can vary, with some markers even having zero effect (Endelman 2011; Gianola et al. 2009; Habier et al. 2011; Hayes and Goddard 2010; Hayes and Goddard 2001; Meher et al. 2022; Meuwissen et al. 2001; Meuwissen et al. 2009; VanRaden 2008). This flexibility allows Bayesian methods to better capture the influence of major alleles at one or a few loci, which is common for traits controlled by a few genes or QTL with larger effects.

From a practical point of view, it is not entirely clear whether it would be better to utilise a MAS or GP based approach to select for traits largely controlled by a single locus. One notable feature of “indirect” single locus marker systems like KASP is that they are best suited to situations with a single target allele (e.g. dominant R genes) that needs to be differentiated from non-target alleles. Additional targets require additional assays. Direct sequence-based systems like PotatoMASH have the advantage of being able to interrogate the allelic structure of a locus directly, meaning that it may be possible to identify or target

multiple alleles. Indeed, the recently updated Potato DArTag V2.0 (Endelman et al. 2024) array includes the ability to interrogate multiple alleles of both *StCDF1* and the *OFP20* gene at the *Ro* locus, making MAS for these traits more practicable. However, variation at these loci doesn't explain the full extent of phenotypic variation, and it might be interesting in the future to explore utilizing combined approaches incorporating locus specific and genome wide information to improve selection for these types of traits.

Insights into the additional information content of haplotypes derived from *haplotype-window*

Examination of the MAF spectrum of the different marker types indicated that the haplotags had a greater ability to resolve all alleles, showing that rare alleles are abundant, and that increasing the detectability of haplotags by utilising the *haplotype-window* approach further increased this resolution. In the tetraploid FRY population, about 1500 sequence-based haplotags identified by *haplotype-window* had a MAF lower than 1%, compared to only around 400 SNP-based haplotags with similar frequencies (Fig. 3). The pattern was more pronounced in the diploid panel, where approximately 3,500 haplotags from SMAP *haplotype-window* had a MAF below 1%, while only about 900 SNP-based haplotags exhibited similar frequencies (Fig. 7).

Although we only tested GBLUP-based models in the diploid panel of this study, we constructed GRMs of both germplasm panels using the different marker datasets to understand their ability to describe genetic relationships within the material. We expected haplotags to provide a more accurate picture of genetic relationships than SNP-based estimates. This expectation is based on two key points: (a) haplotags capture a more detailed allelic variation at loci, reflecting the actual allelic composition of the allelic variation at a locus compared to single SNPs, and (b) SNPs are often selected based on their moderate to high MAF, which means they typically represent older mutations. New mutations are often at low frequency and can be lost before they reach a detectable level in the population (Meuwissen et al. 2014). As a result, SNP-based GRMs primarily reflect older relationships from more distant ancestors, whilst haplotype-based GRMs will be better at tracing recent genetic changes and therefore provide a more accurate representation of genetic relationships than SNP-based GRMs.

In the tetraploids, the most striking result was that PotatoMASH-derived data was very comparable to the GBS derived data in being able to infer relatedness despite relying on over 30 times fewer loci. However, in general, the GRMs of the haplotags do not show more pronounced genetic structure compared to the SNP-based GRMs. This could be due to the small genetic pool of the panel, which consists of individuals derived from the same breeding programme, or because many of the haplotags were too rare to provide meaningful contributions to the genetic relationships. Although haplotags are expected to capture more allelic diversity information than SNPs, the high frequency of rare haplotags likely limits their potential additional benefit.

When estimating genomic relationships for the more diverse diploid panel, all three PotatoMASH-derived marker sets proved effective in capturing genomic relationships and population structure, as demonstrated by the GRMs and PCA based on the GRMs (Fig. 7 and 8). A distinct pattern of sub-population structure was observed, with the Meijer materials showing genetic deviation from the other genepools, as previously reported by Vexler et al. (2024). Notably, in this more diverse diploid panel, the haplotags performed better than the SNPs in capturing genetic relationships, as shown by more accurate grouping of the Meijer materials and higher variance explained in the first principal component (PC) of the haplotags PCA compared to the SNPs PCA.

Conclusions

PotatoMASH was originally conceived as a low-cost marker system to facilitate the combined application of MAS and GS in potato breeding. Given the specific breeding contexts in which we envisioned its use, we adopted an approach that minimizes the number of loci surveyed, ultimately selecting 339 loci based on observed LD and recombination patterns in potato (Leyva-Pérez et al. 2022). In this study, we have demonstrated the potential for GS and explored the utility of haplotags that can be derived from the targeted amplicons to increase the prediction accuracy using the platform, laying the ground for this application in a real-world breeding scenario. It’s worth noting that “low to medium-density” genotyping based on targeted amplicon sequencing is becoming increasingly cost-effective and has been accompanied by the advent of publicly available tools such as the Potato DArTag array(Endelman et al. 2024), with more systems potentially in development. Insights gained into the use of short read haplotyping for GS in this study may be useful for the application of similar systems as they become more common.

Acknowledgements

We thank the members of the public-private partnership "A new method for potato breeding: the 'Fixation-Restitution' approach" and SusCrop ERANET funded project "DIFFUGAT: Diploid Inbreds For Fixation, and Unreduced Gametes for Tetraploidy" (Averis Seeds B.V., Bejo Zaden B.V., Danespo A/S, Germicopa, Den Hartigh B.V., SaKa Pflanzenzucht GmbH & Co. KG, Meijer Potato, and Teagasc) for providing their support.

This research was carried out using the Teagasc high-performance computing cluster and storage systems, and the support of Dr. Paul Cormican is greatly appreciated.

We acknowledge the full support of the Teagasc Potato Breeding Programme.

Author contribution statements

L.V., S.B. and D.M. conceived and designed the study, L.V. performed formal bioinformatics, statistical and GP analysis, data curation, investigation, methodology and software. D.M., S.B., A.K and J.K. advised on statistical and GP analysis, methodology and visualization. M.d.l.O.L.-P. and T.R. advised on bioinformatics processing. T.R. provided the SMAP software. D.M., H.J.v.E., and D.G. and R.G.F.V. obtained resources. D.M., H.v.E. and R.G.F.V. supervised the research. L.V., S.B. and D.M. wrote the original draft. L.V., S.B., A.K., T.R., M.d.l.O.L.-P., R.G.F.V. and D.M edited the manuscript. All authors reviewed and agreed to the published version of the manuscript.

Supplementary files

Suppl. File 1: GFF file for SMAP *haplotype-window*, with loci coordinates of PotatoMASH regions in the reference genome DM_v6.1

Suppl. File 2: Phenotypic and PotatoMASH genotypic data for 607 individuals in the tetraploid population.

Suppl. File 3: Prediction Accuracy (PA) for 'Off-the-Field' fry colour in the tetraploid population, using different marker sets, training-testing set combinations, and prediction algorithms.

Suppl. File 4: Phenotypic and PotatoMASH genotypic data for 558 individuals in the diploid panel.

Suppl. File 5: Prediction Accuracy (PA) results for 23 traits in the diploid panel, analysed across three PotatoMASH marker sets and seven prediction algorithms.

Suppl. File 6: Effect of different marker sets, SNPs, SNP-based haplotags from SMAP *haplotype-sites* and sequence-based haplotags from SMAP *haplotype-window* on the mean PA for 23 agronomic and quality traits in diploid potato. The Mean PA is calculated as the average Pearson correlation obtained in a k = 5-fold cross-validation scenario, repeated over 10 iterations. The results are presented with a unified y-axis (ranging from 0 to 1) for all traits, providing a comprehensive overview.

Chapter 6

General Discussion

Over 20 years ago, Peleman and Rouppe van der Voort (2003) described “Breeding by Design” as “a concept that aims to control all allelic variation for all genes of agronomic importance”, suggesting that it could “be achieved through a combination of precise genetic mapping, high-resolution chromosome haplotyping and extensive phenotyping”. To a large extent, this has remained the vision for plant breeding across all species for the 20+ years since the authors first talked about it. It could be argued that, despite the optimistic outlook of the authors at the time, Breeding by Design still hasn’t materialised in its fullest form, and whilst it is probably much closer in some crops, in others, such as potato, the journey towards it has been more difficult. At the time, the authors had limited awareness of just how many haplotypes could exist per gene, the challenge of tracking them with specific markers, or the complexity of linking each to its phenotypic effect.

As I presented in the introduction to this thesis, the small progress towards Breeding by Design is partly due to the tetraploid, outbreeding nature of cultivated potato, which makes the organised progress towards collecting “optimal” sets of alleles much more difficult. In addition to this, breeding by design has suffered a practical limitation in the ability to control “all allelic variation for all genes” due to the high cost of deployment of the types of genome wide marker systems that are necessary for its implementation. Marker-assisted selection (MAS), genome-wide association studies (GWAS), and genomic selection (GS) provide the foundation for modern genomics-assisted breeding strategies, but their effectiveness depends on the accuracy, scalability, and affordability of genotyping technologies. As outlined in Chapter 1 (General Introduction), advances in sequencing technologies have expanded the range of available genotyping tools, offering potential improvements in genomic-assisted breeding. However, many of these tools are either highly specialized for specific applications or too expensive for routine use in breeding programs. This creates a gap where breeders need cost-effective, high-resolution genotyping solutions that can support a range of breeding applications, including identifying marker-trait associations, tracking heterozygosity during inbreeding, improving genomic prediction (GP) models, and optimizing selection strategies.

This thesis was driven in part by the need to fill this gap. The development of PotatoMASH initially started as a thought experiment, with the goal of determining how inexpensive a genotyping platform could be while still remaining effective for all major breeding applications. It became apparent that a process called GT-Seq (Genotyping in Thousands by Sequencing) (Campbell et al. 2015), developed originally in the fish trout, had some potential in the area because of its low cost and relative technical achievability at the level possessed by many medium-sized breeding programmes. However, the relatively low number of loci targeted by the approach (<500) caused concern.

In addition to the potential to develop a low cost marker platform, this thesis coincided with the inception of Fixation Restitution (Fix-Res) breeding. Whilst a low cost genome wide marker system would have applicability across all potato breeding modalities, a key component of Fix-Res breeding is the integration of genomic tools in a breeding-by-design

framework, enabling breeders to facilitate selection, track recombination, and monitor genetic diversity. As a consequence, much of the work done to demonstrate the applicability of utilising PotatoMASH was performed in the context of Fix-Res breeding, but it's important to highlight that most of the lessons learned during the thesis are generalizable to the use of haplotag-oriented amplicon sequencing approaches in any potato breeding modality.

This thesis makes the following contributions in this research area:

- **Introduction of a novel low-cost genome-wide scanning marker platform** capable of yielding short read-backed haplotypes (haplotags) that have potential to increase the multi-allelic information content of the assay.
- **Identification of QTL for multiple traits** under selection in Fix-Res diploid founder material and establishing the concept that SNPs and haplotags have a complementary ability to detect QTL, increasing QTL discovery when both are used.
- **Development of genomic prediction models** in Fix-Res diploid material founder material and comparison of read-backed haplotypes with bi-allelic SNPs.
- **Identification of genomic regions recalcitrant to inbreeding** in the Fix-Res breeding pool, and a demonstration of how **haplotags can offer advantages over SNPs in tracking the zygosity status** of material in diploid potato breeding.

The goal of this final Chapter is to highlight the novel contributions of this thesis. It provides a synthetic overview of how these findings fit together and contribute to a broader system of genome-based breeding, in line with the Breeding by Design concept. It also outlines how these advances can help advance the state of the art in breeding.

Haplotype-based genotyping in the context of modern breeding approaches

Typically, high-density SNP panels have been utilised to capture QTL effects in association and GP studies. For the successful implementation of GWAS and GS, high-density SNP panels have been traditionally considered necessary to capture genetic variation across the genome, ensuring that all QTL are in LD with at least one marker, thereby capturing a significant portion of the genetic variance (Heffner et al. 2009; Slater et al. 2016; Vos et al. 2017). In fact, most GWAS and GS studies in potatoes rely on GBS data containing tens to hundreds of thousands of SNPs, such as the 186k SNPs (Sverrisdóttir et al. 2017), 46k SNPs (Byrne et al. 2020), 39k SNPs (Wilson et al. 2021), or SNP arrays such as SolSTW 20K Infinium array (Prodhomme et al. 2020; Vos et al. 2022), and the 8.3k SNPs from the SolCAP potato genotyping array (Rosyara et al. 2016; Stich and Van Inghelandt 2018).

However, the high cost of dense SNP genotyping presents a major challenge for large-scale breeding programs. Smaller and more cost-effective marker sets have the potential to serve these purposes. To achieve this, several strategies have been developed to reduce the number of SNPs used in GWAS and GS while maintaining accuracy. These approaches include LD pruning, which selects markers based on LD between adjacent SNPs (Selga et al. 2021a), prioritizing trait-associated markers identified in GWAS (Byrne et al. 2020), and

feature selection techniques to identify the most informative SNP subsets (Aalborg et al. 2024). While these methods successfully enable marker-trait associations and GP, they often lead to reduced statistical power or lower predictive accuracy compared to high-density genome-wide markers. The ongoing challenge remains in balancing genotyping costs with sufficient allelic information to ensure reliable GWAS discovery and GS performance.

The challenge lies in reducing the number of markers to minimize genotyping costs and data processing/storage requirements while simultaneously maximizing allelic information to support the routine implementation of these genomic-based approaches in commercial breeding programmes. In this thesis, we sought to achieve this by using read-backed haplotypes which can capture the full complexity of allelic variation more effectively than traditional bi-allelic SNPs. This is particularly advantageous for crops like potato, which exhibit high heterozygosity and complex polyploidy inheritance. Previous studies have reported high nucleotide diversity between haplotype genomes, with estimates of 1 in 50 nucleotides. In non-coding and coding regions, there are SNPs across haplotypes at a frequency ranging from 1 in 25 to 1 in 15, respectively (Uitdewilligen et al. 2013). This high level of polymorphism is reflected by a high number of different haplotypes, ranging between 5 and 20, and a large nucleotide diversity between them, depending on the genomic region and methodology used (Uitdewilligen et al. 2013; Wolters et al. 2010; Uitdewilligen et al. 2022).

PotatoMASH was developed as a haplotype-based genotyping tool, capable of simultaneously detecting both SNPs and short read-backed haplotypes (haplotags) when combined with appropriate analytical tools such as SMAP. At the outset of this research, we hypothesized that haplotype-based markers might outperform traditional SNPs in quantitative genetics applications; based on the assumption that haplotags would be in stronger linkage disequilibrium (LD) with specific QTL alleles than individual SNP alleles (Calus et al. 2008; Hess et al. 2017; Meuwissen et al. 2014). As demonstrated in this thesis, PotatoMASH effectively captures multi-allelic haplotype variation, providing greater genetic resolution than single SNP markers. Across the analysed breeding material, an average of 6 haplotype alleles (haplotags) per locus was detected in the tetraploid panel, with individual loci exhibiting 2 to 14 haplotags. In the more diverse diploid panel, the number of haplotypes was even higher, with an average of 8–9 haplotags per locus, ranging from 2 to 30 haplotypes per locus. While this may or may not capture the complete extent of allelic variation at each locus, they are consistent with the previous studies mentioned above, which reported similarly high numbers of distinct alleles in potato germplasm.

Therefore, PotatoMASH appeared useful in various breeding applications, providing new insights into how haplotype-based selection can improve breeding efficiency. Although this research was conducted within the Fix-Res framework, these findings must be considered in the broader context of other breeding strategies, including tetraploid breeding, diploid F1 hybrid breeding, and progressive heterosis, all of which share key breeding phases: (1) pre-breeding activities, (2) introgression of beneficial loci, and (3) clonal selection following

crossing or selfing. A key advantage of a system like PotatoMASH is that it provides a unified marker system that can facilitate all of these genomic-based breeding activities across different strategies, ensuring a consistent and standardized genotyping approach that produces a single data type for use in multiple applications and stages in a breeding programme. At the same time, PotatoMASH can still yield “traditional” SNP-based data remain highly compatible with existing genomic resources, allowing breeders to integrate PotatoMASH data with prior SNP-based studies or existing SNP arrays. By combining haplotypes and SNPs within a single system, PotatoMASH maximizes flexibility in GS and trait mapping, enabling breeders to adapt their selection strategies to the specific needs of each breeding program.

Genome-based breeding; what strategy to adopt, and when?

Modern breeding strategies often integrate GWAS, MAS, and GS as complementary tools applied across different phases of the breeding cycle. As outlined in the general introduction, GWAS (including bi-pawrental mapping and other marker-trait association approaches) is commonly used to identify loci associated with important traits by analysing genotype-phenotype relationships in diverse populations. The resulting markers can be deployed in MAS to rapidly select individuals carrying favourable alleles. However, MAS is most effective when traits are controlled by major genes or large-effect QTL, such as disease resistance, where a single marker can reliably indicate trait presence

Improving complex traits in potato, such as yield, dry matter content, chipping colour, and dormancy, is more challenging due to their polygenic nature and strong genotype-environment interactions (Ewing and Struik 1992; Hu et al. 2023; Leonel et al. 2017; Navarro et al. 2011). These traits are typically influenced by numerous small-effect loci, often interacting through regulatory networks and epistasis, and further complicated by extended linkage disequilibrium (Hill et al. 2008; Kloosterman et al. 2010; Stich and Gebhardt 2011). Additionally, QTL for multiple traits frequently co-localize, as shown in several studies (Acharjee et al. 2018; Kloosterman et al. 2013; Sharma et al. 2018) and in our own results (Chapter 3), making it difficult to isolate causal variants and limiting the broader applicability of diagnostic markers across populations.

While MAS, as practiced in most breeding programmes, is currently based on the use of single bi-allelic markers diagnosing the presence or absence of a target allele, this approach is poorly suited to such complex traits. Interestingly, if genome-wide marker systems previously used to characterize QTL could also inform selection, MAS for complex traits might become more feasible. In particular, since complex traits may be determined not only by multiple loci (genes) but multiple alleles of these loci, a system capable of interrogating the underlying allelic variation of individual loci would be very useful. Read-backed haplotyping approaches like PotatoMASH offer this feature.

GWAS

To explore the above concept, we performed a multi-trait GWAS using both SNPs and read-backed haplotypes (haplotags) generated through PotatoMASH. To our knowledge, this study represents the most extensive multi-trait GWAS conducted in diploid potato to date, which encompassed 23 agronomic and quality traits across 558 genotypes from six European breeding programs. Due to the diverse genetic backgrounds represented, the panel exhibited strong population structure, which we accounted for in downstream analyses to detect population-specific associations. Moreover, it is the first GWAS in potato to use short read-backed haplotypes as markers, providing a novel perspective on multi-allelic marker utility in this crop. This allowed for an unprecedented resolution in detecting QTL for complex traits. The study identified a total of 37 QTL, with only 10 (27%) QTL consistently detected using both SNPs and haplotags. Haplotags alone identified 14 (38%) additional QTL that were undetectable with SNPs, highlighting their ability to capture recombination events and multi-allelic variation that bi-allelic SNP markers fail to detect. Conversely, 13 (35%) QTL were uniquely detected by SNPs, an unexpected finding given that haplotags theoretically provide superior genomic resolution.

Upon an investigation into the underlying genetic architecture of these QTL, we discovered key differences in how SNPs and haplotags capture marker-trait associations. For QTL identified solely by haplotags, my analysis revealed that the significant haplotags often contained individual SNPs that were also present in other non-significant haplotags, demonstrating that haplotags provided a more comprehensive representation of genetic variance at these loci. Conversely, for QTL detected only with SNP data, we found that the significant SNPs were scattered across multiple haplotags, each occurring at a low allele frequency in the population. This reduced the statistical power to detect associations using haplotags alone. These findings suggest that increasing population size in future studies could enhance statistical power, allowing more haplotag-QTL associations to be detected. We concluded that the complementary nature of SNPs and haplotags indicates that both marker types should be used in parallel to maximize QTL detection power and ensure a more comprehensive understanding of the genetic basis of complex traits. Examples of these scenarios are further explored in Chapter 3.

To account for the strong population structure observed in the panel, we repeated the GWAS analysis separately for two main sub-populations, each representing different breeding programs. This analysis uncovered unique genetic associations that were not detected when analysing the full panel, underscoring the importance of accounting for population structure in association studies. Of the 37 identified QTL, 15 (41%) were found to be population-specific, with eight detected using haplotags and seven using SNPs. While analysing smaller sub-populations typically reduces statistical power, particularly for rare haplotypes, working within a genetically closer pool can reduce background noise and improve detection of common alleles. This further demonstrates that SNPs and haplotags capture different signals and perform best in complementary contexts. Beyond validating known QTL, such

as those for Tuber Shape and Yellow Flesh Colour, which confirmed the reliability of our approach, a major outcome of this study was the discovery of 19 novel QTL across eight chromosomes, contributing valuable genetic insights for potato breeding.

Genomic Prediction

Given the limited number of strong marker-trait associations identified through GWAS, we evaluated an alternative approach to enable selection for complex traits in breeding programs: GP using low-density marker sets (Chapter 5). We tested the effectiveness of PotatoMASH-derived markers by comparing their performance to high-density GBS data in a tetraploid population phenotyped for fry colour. PotatoMASH generated 2,236 SNPs, 2,000 SNP-based haplotags, and 3,390 direct-read haplotags, compared to 43.6k SNPs from GBS. Despite this substantial reduction in marker density, the decline in predictive accuracy (PA) was only 14% with PotatoMASH SNP data and 9% with haplotag data, relative to GBS SNPs. This modest drop in PA underscores the potential of amplicon sequencing as a cost-effective genotyping tool for large-scale GS without significantly compromising accuracy.

In the diploid panel, we developed GP models for the 23 agronomic and quality traits. As with the GWAS, this study is the most extensive GP analyses in diploid potato, and uniquely applies read-backed haplotypes within a low-density, cost-effective genotyping system. We achieved moderate to high predictive accuracy across all traits, including Tuber Regularity, Skin Brightness, and Presentability of Tubers, for which no significant QTL were detected in the GWAS. Across the 23 traits, haplotags outperformed SNPs in 11 traits (48%), including Canopy Stage 1, Canopy Stage 2, Tuber Length, Total Tuber Number, Tuber Regularity, Yellow Skin Colour, Presentability of Tubers, Sensitivity to Common Scab, Enzymatic Browning, Dry Matter Content, and Sprout Dormancy. In contrast, SNPs obtained higher predictive accuracy for six traits (26%), including Yield, Skin Smoothness, Cooking Type, After-Cooking Blackening, Chipping Colour 1_8, and Chipping Colour 2_8. The remaining six traits (26%) produced same or mixed results across different prediction models. These results align with our GWAS findings, where SNPs and haplotags demonstrated different advantages in detecting marker-trait associations, further reinforcing their complementary roles.

Interestingly, for traits controlled by major QTL, such as Tuber Shape, Yellow Flesh Colour, and Maturity, predictive accuracy was nearly identical between SNPs and haplotags, suggesting that the GP model had a greater influence on accuracy than marker type. In these cases, Bayesian models (BayesA, BayesB, BayesC) outperformed GBLUP and ridge regression methods, likely because they assign greater weight to large-effect loci. Conversely, for complex polygenic traits such as Dry Matter Content, Tuber Regularity, and Enzymatic Browning, the marker type had a stronger influence on predictive accuracy. This suggests that when multiple small-effect loci control a trait, the ability to detect them is more sensitive to the marker type used rather than the model chosen, as allele frequency and the

ability to capture multi-allelic variation become more relevant in small effect loci of genetically complex traits prediction.

In the GP study, haplotypes were derived using two approaches: 1) SNP-based haplotags, generated with the SMAP haplotype-sites module, which phases alleles based on predefined SNP positions (sites). 2) Direct read-mapping haplotags, generated with the SMAP haplotype-window module, which reconstructs haplotypes directly from sequenced reads without relying on prior SNP calls. The key difference between these methods lies in how haplotypes are defined and constructed. The haplotype-sites approach depends on pre-called SNPs, while the haplotype-window approach identifies haplotypes independently, capturing additional genetic variation, including indels. Among the 11 traits where haplotags outperformed SNPs, six exhibited the highest PA with haplotags derived from haplotype-window, including Canopy Stage 1, Total Tuber Number, Tuber Regularity, Yellow Skin Colour, Presentability of Tubers, and Dry Matter Content. Two traits, Tuber Length and Sprout Dormancy, achieved higher PA with SNP-based haplotags, while the remaining three traits showed similar PA across both haplotyping approaches.

These findings indicate that, at the scale tested in this study, there is no substantial difference in performance between SNPs and haplotags for GP. Either marker type can be used, with the choice depending on which provides higher predictive accuracy for a given trait. This flexibility underscores the utility of amplicon sequencing-based platforms like PotatoMASH, which enable simultaneous detection of both marker types at low cost. The high read depth and accurate dosage calling offered by amplicon sequencing ensures reliable genotyping, making GS more accessible and scalable for large-scale breeding programs.

Relevance for breeding and practical implications

The integration of GWAS, MAS, and GS is an important goal in modern tetraploid and diploid F_1 hybrid potato breeding particularly in enhancing trait discovery, selection efficiency, and genetic gain. Each of these genomic tools contributes uniquely to the breeding process. GWAS plays a central role in identifying trait-associated markers, enabling the discovery of genetic loci linked to key agronomic and quality traits. MAS is particularly effective for selecting major-effect QTL early in the breeding process, such as disease resistance genes, allowing breeders to eliminate inferior genotypes before extensive field trials. Meanwhile, GS is invaluable for predicting polygenic traits, improving the selection of parental lines, and ranking candidates based on their genomic estimated breeding values (GEBVs), which enhances the efficiency of both early- and late-stage selection.

In any breeding strategy, the pre-breeding phase is indispensable for establishing genetic foundations before large-scale selection begins. In this phase GWAS allows the identification of trait-associated QTL in genetically diverse populations, enabling the discovery of markers linked to important agronomic and quality traits. Once these markers are identified, MAS is applied to introgress and fix major-effect QTL, particularly for disease resistance and other one of few major-effects QTL traits. Simultaneously, GS aids in the

selection of parental lines by predicting their breeding values, optimizing the choice of crosses and improving genetic gain in subsequent generations.

Tetraploid breeding is characterized by an intensive early selection phase, when large numbers of seedlings must be evaluated, which is currently based primarily on phenotypic performance to eliminate poorly performing clones. By the fourth year of the breeding cycle, when the number of seedlings is reduced from tens of thousands to a few thousand, MAS is applied to select for major-effect QTL, such as disease resistance. At present, GS is not routinely applied in early selection (or at any stage) due to the high cost of genome-wide genotyping for large populations, but its use has been suggested for later selection stages when the breeding population size has already been narrowed down. However, with the advent of cost-effective amplicon sequencing-based genotyping tools such as DArTag and PotatoMASH, MAS and GS can now be applied in a single genotyping step. This advancement makes GS feasible at early selection stages, significantly improving efficiency by ranking candidates based on GEBVs from the outset.

As breeding progresses to later stages, fewer clones undergo extensive field trials, where phenotypic data collection becomes more detailed. This additional phenotypic information refines the GP models, increasing their prediction accuracy over successive selection cycles. Over time, as training datasets expand, GS can be applied to a broader range of traits, reducing the need for expensive and labour-intensive field trials while accelerating genetic gain.

In diploid F_1 hybrid breeding, breeding efforts focus on developing elite inbred parental lines to ensure strong hybrid performance. Once superior parental lines are selected, MAS enables the introgression and fixation of beneficial alleles, such as resistance genes or key traits like self-compatibility (*Sli*), which is fundamental for allowing selfing in inbred lines before hybridization. GS further enhances parental selection by predicting their breeding value, ensuring that only the most promising genotypes are used for hybrid crosses.

A major advantage of amplicon sequencing-based genotyping is its ability to reduce genotyping costs while maintaining high accuracy. The combination of targeted sequencing with read-backed haplotyping offers a genome-wide, cost-effective alternative to traditional high-density SNP arrays and GBS. Unlike conventional approaches, where different genotyping platforms are used at different breeding stages, such as KASP for MAS and SNP arrays or GBS for GWAS, and GS, targeted amplicon sequencing provides a unified platform. With this approach, genotyping is performed once per clone, and all genomic-based breeding activities (GWAS, MAS, GS) can be conducted using the same marker dataset throughout the entire breeding cycle.

This “single-genotyping event” strategy significantly reduces the need for repeated genotyping across different breeding phases, cutting down on costs and logistical complexity. By integrating GWAS for marker discovery, MAS for early selection, and GS for multi-trait prediction within a single, cost-efficient genotyping platform, breeders can

streamline selection, improve genetic gain per cycle, and make GS more accessible and scalable for large-scale potato breeding programs.

Comparison of PotatoMASH with Current Genotyping Tools

To contextualize the role of PotatoMASH in modern breeding pipelines, we compare its core features to other genotyping platforms commonly used in potato (Table 1). A detailed overview of existing technologies is provided in the general introduction (Chapter 1).

KASP assays (https://excellenceinbreeding.org/) offer a cost-effective solution for targeted marker genotyping but they lack genome-wide coverage and are not optimal for broader genomic-based applications. SNP arrays offer high-density genotyping but are limited by fixed panels and ascertainment bias (Vos et al. 2015). GBS (Elshire et al. 2011) provides genome-wide coverage but is costly and prone to missing data and uneven coverage, especially problematic in polyploids .

Amplicon-based genotyping platforms in potato, such as DArTag (www.diversityarrays.com) and Flex-Seq (www.biosearchtech.com), have been developed in parallel to PotatoMASH as cost-effective alternatives. These systems target specific loci through multiplexed PCR, enabling consistent read depth, reduced missing data, and accurate dosage calling. However, they primarily deliver bi-allelic SNPs as the standard output.

Although PotatoMASH targets only 339 loci, it consistently delivered high-resolution data across the diverse populations analysed in this thesis, with 2 to 38 haplotags per locus depending on genetic diversity and analytical method. This multi-allelic output allows comprehensive capture of genetic variation, including rare alleles and indels that may be overlooked by bi-allelic SNP-based platforms.

By contrast, commercial amplicon-based platforms such as DArTag V2 (Endelman et al. 2024) offer a fixed panel of approximately 4,000 loci, providing broader marker density but less flexibility in design. While DArTag and Flex-Seq typically do not deliver haplotype calls directly, raw sequencing data can be extracted and re-analysed with tools such as SMAP to derive multi-allelic haplotypes.

In terms of cost, PotatoMASH was estimated at €4–5 per sample (excluding labour) in Chapter 2, while commercial platforms such as DArTag and Flex-Seq typically cost around €18 per sample. However, when considering overhead, data processing, and management, the total costs may converge. Overall, DArTag and PotatoMASH offer comparable capabilities in terms of marker type, dosage accuracy, and suitability for breeding applications.

Overall, DArTag and PotatoMASH offer comparable capabilities in terms of marker type, dosage accuracy, and suitability for breeding applications. However, PotatoMASH offers greater flexibility in assay design and data access, making it particularly attractive for programs with in-house technical capacity. The choice between them depends largely on practical factors, such as control requirements, cost structure, and available expertise.

Table 1. A comparison of key genotyping technologies, including PotatoMASH, SNP arrays, KASP, GBS, and amplicon-based sequencing (DArTag, Flex-Seq), based on marker type, flexibility, cost, suitability for breeding applications, key advantages, and limitations. This table highlights the strengths and constraints of each platform.

Genotyping Tool	Marker Type	Flexibility	Cost per Sample	Suitability for Breeding	Key Advantages	Key Limitations
PotatoMASH	SNPs & short-read haplotypes (haplotags)	**High** (new loci can be added)	**Low**	**High** (QTL discovery, GS, MAS, tracking heterozygosity)	Captures SNPs & haplotags simultaneously, cost-effective, high flexibility, high-throughput.	339 loci, requires lengthy library construction for in-house use, not yet commercialized
SNP Arrays	Fixed SNPs	**Low** (fixed marker panels)	**Medium to High**	**Moderate** (QTL discovery, MAS, GS)	High-density genotyping, widely used	Ascertainment bias, expensive, fixed panels
KASP	Targeted SNPs	**Low** (fixed SNP assays)	**Very Low**	**Low** (diagnostic marker screening)	Low-cost, fast, high-throughput, useful for trait-specific markers.	Limited to known SNPs, not suitable for broader breeding activities
GBS	Random SNPs	**High** (random genome-wide SNPs)	**High**	**Moderate** (genome-wide discovery, QTL discovery, GS)	Genome-wide coverage, good for novel marker discovery	High missing data, requires imputation, costly for commercial use
Amplicon-based Genotyping (e.g. DArTag, Flex-seq)	Targeted SNPs	**High** (predefined target SNPs loci can be added)	**Low**	**High** (QTL discovery, GS, MAS)	Currently commercially available service, cost-effective, high-throughput.	Limited flexibility: adding loci requires collaboration with service provider

Genomic tools in Fix-Res: the role of amplicon sequencing and read-backed haplotyping in breeding and selection

The Fixation-Restitution (Fix-Res) breeding strategy provides a structured approach to genetic improvement by combining incremental allele fixation at the diploid level with high heterozygosity restitution at the tetraploid level. The successful implementation of this strategy relies on accurate genetic tracking, recombination monitoring, and selection for beneficial alleles across breeding generations (Figure 1). To achieve this, Fix-Res integrates MAS, GWAS, and GS, all of which require a cost-effective genotyping platform capable of supporting multiple breeding applications.

Most of practical applications of PotatoMASH as a breeding tool in this thesis were developed within the framework of Fix-Res to address those needs.

In addition to GWAS, MAS and GS, another key aspect of tracking the genetic composition of clones throughout the Fix-Res breeding process is tracking inbreeding. Chapter 4 explored the use of PotatoMASH for monitoring inbreeding. By analysing inbreeding levels in the Wageningen diploid breeding program and across three generations of selfing, haplotag-based measurements proved to be a reliable tool for assessing homozygosity and heterozygosity, even in the absence of parental genomic information. This highlights the value of haplotype-based genotyping for detecting genome-wide homozygosity changes and inferring selfing rates in breeding materials across Fix-Res cycles.

Overall, with a single round of genotyping per new clone, PotatoMASH enables breeders to integrate all of these genomic tools and make data-driven decisions at every stage of Fix-Res breeding. The following sections will discuss how PotatoMASH could be applied at each stage of Fix-Res (Figure 1).

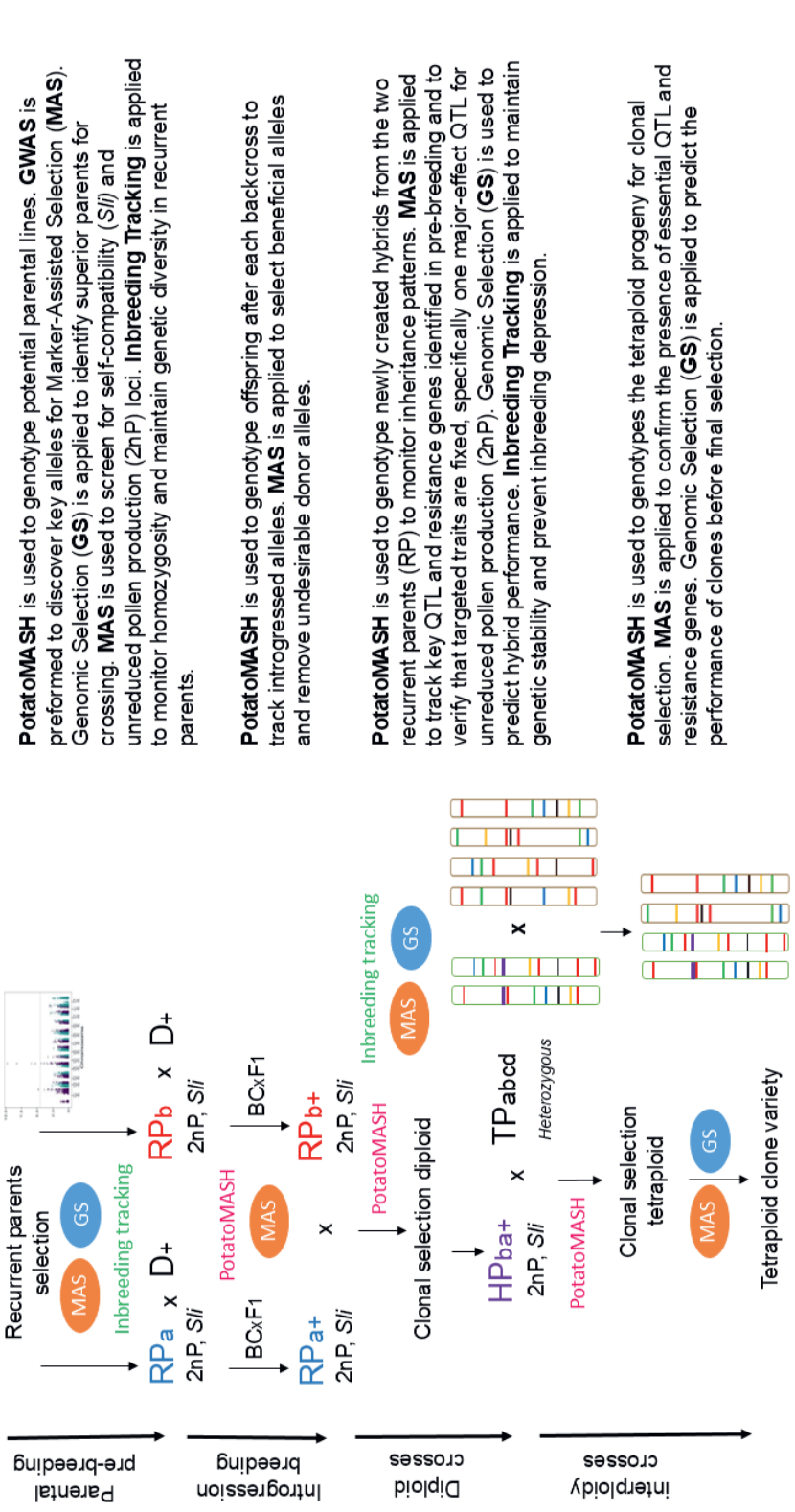


Figure 1. Overview of the Fixation-Restitution (Fix-Res) Breeding Strategy. This scheme illustrates the sequential integration of Marker-Assisted Selection (MAS), genome-wide association studies (GWAS), Inbreeding Tracking, and Genomic Selection (GS) using PotatoMASH to support decision-making across different breeding

phases. Recurrent parents (RP) are genotyped during pre-breeding to discover alleles via GWAS and to assess genetic diversity. Donor lines (D) are introduced to incorporate traits of interest, and MAS is used to track and recover desirable alleles through successive backcrosses. Two improved recurrent parents (RP) are then crossed to generate a hybrid progenitor (HP), which is evaluated for key loci before being used in interploidy crosses with tetraploid breeding material. GS is applied at multiple stages to guide parent selection and to predict clone performance. Inbreeding Tracking supports genetic stability by monitoring homozygosity across cycles. A single round of genotyping per clone using PotatoMASH supports all genomic applications throughout this process. (Figure adapted from Clot (2023).*

Application of genomic tools across Fix-Res breeding phases

Fix-Res consists of multiple breeding phases, each requiring specific genomic tools for selection and genetic tracking. The integration of PotatoMASH at each stage ensures that allele inheritance is monitored, genetic gain is maximized, and inbreeding is controlled.

Pre-breeding: parental selection and trait discovery

The pre-breeding phase of Fix-Res aims to establish a genetically diverse and self-compatible (*Sli*) recurrent parent (RP) population, carrying favourable agronomic traits and the ability to produce unreduced (2n) pollen. This phase is foundational, as it sets the genetic foundation for all subsequent Fix-Res breeding steps (Figure 1). The success of pre-breeding depends on accurate identification of trait-associated loci through GWAS and the targeted selection of key alleles using MAS.

A key component of this stage is genome-wide marker discovery, which is used for trait mapping and parental selection. GWAS enables the identification of QTL associated with key agronomic traits, and MAS ensures the selection of lines carrying favourable alleles. Haplotype-based genotyping further enhances this process by tracking recombination events and identifying complex genetic patterns that bi-allelic SNPs may not capture.

A key requirement in the pre-breeding phase is the selection of parental lines that carry self-compatibility (*Sli*) loci. This can be done using MAS with two diagnostic markers developed by Clot et al. (2020). Currently, these markers are designed for KASP assays, but the *Sli* locus on chromosome 12 is not covered by the core PotatoMASH loci. Given its importance, this locus should be added to enhance the platform's ability to screen for self-compatibility in Fix-Res breeding.

A compelling example of pre-breeding facilitated by PotatoMASH is the identification of two major QTL for unreduced pollen (2nP) production (Clot et al. 2024), a key trait in Fix-Res breeding that ensures chromosome restitution in interploidy crosses. The study by Clot et al. (2024) investigated the mechanism and inheritance of unreduced pollen in three biparental diploid populations, which were genotyped using PotatoMASH. This analysis identified major-effect QTL for 2nP production on chromosomes 7 and 12, which co-localized with *StJR1* and *StJR2*, respectively. These genes are homologous to *AtJAS*, a key regulator of meiotic restitution and 2n gamete formation. PotatoMASH-derived haplotags enabled the identification of haplotypes associated with the functionality of these candidate genes, allowing breeders to screen and select Fix-Res recurrent parents that carry *StJR1* and *StJR2* haplotypes linked to high 2nP production. These findings reinforce the genetic

foundation necessary for efficient interploidy breeding and chromosome restitution in later Fix-Res stages.

Monitoring heterozygosity in recurrent parental lines is an important aspect of Fix-Res pre-breeding, ensuring that beneficial alleles are retained while avoiding excessive inbreeding depression. Inbreeding through selfing increases homozygosity, yet certain genomic regions consistently maintain heterozygosity, likely due to selective pressures preserving beneficial alleles. Clot in his PhD thesis (2023) demonstrated that recurrent parents in Fix-Res breeding do not necessarily need to be fully inbred but can remain partially heterozygous, maintaining genetic diversity vital for agronomic performance.

Introgression breeding: tracking alleles during backcrossing and selection

Once a strong recurrent parent (RP) population is established, the next phase involves introgressing novel alleles from donor parents (D^+) into the recurrent parent population. This process is essential for introducing new resistance genes and desirable traits. Successive backcrosses then eliminate unwanted donor background while retaining key introgressed loci. Genotyping is applied after each backcross to track selected alleles and monitor recombination patterns, ensuring that beneficial genetic material is preserved. MAS plays a key role in this phase, allowing breeders to efficiently select for favourable alleles while eliminating undesirable donor alleles, ultimately leading to the development of improved recurrent parents (RP^+) for further breeding.

Diploid crosses: selecting the best hybrid progenitors

In the diploid breeding phase, two improved recurrent parents (RP^+) are crossed to generate hybrid progenitors (HP), which undergo clonal selection before advancing to interploidy breeding. The objective is to combine favourable alleles from both parents while maintaining genetic diversity and stability. PotatoMASH is applied to genotype all newly created hybrids, allowing breeders to track inheritance patterns and recombination events. This ensures that beneficial alleles identified in pre-breeding (e.g., *Sli*, 2nP, resistance genes) are retained, while undesirable variants are monitored and eliminated through selection.

MAS is used to track and maintain key QTL and resistance genes, ensuring that targeted traits are successfully passed to the next generation. In particular, MAS is applied to verify that hybrid progenitors are fixed for *Strj1* and heterozygous for *Strj2* or vice versa for unreduced pollen production (2nP) (Clot et al. 2024), a key trait for successful chromosome restitution in interploidy crosses. In the work presented in Chapter 3, we identified a wide array of QTL associated with key agronomic and quality traits in diploid potato populations, which can be leveraged to maintain beneficial alleles through marker-assisted selection.

Using the same PotatoMASH-derived genotyping data, GS is applied in the clonal selection of hybrid progenitors, using PotatoMASH-derived genomic data to predict the breeding value of hybrid progenitors. This enables the selection of promising hybrid progenitors based on predicted performance, complementing MAS by capturing polygenic effects that may not be detected through QTL-based selection. In Chapter 5, we demonstrated the utility of genomic

prediction models, based on the same PotatoMASH data as used in the QTL discovery of Chapter 3, to estimate breeding values for a broad range of complex traits in diploid material.

Inbreeding tracking is also performed in this phase, evaluating homozygosity levels in hybrid progenitors to stabilize beneficial alleles while preventing excessive inbreeding depression. This process ensures that Fix-Res hybrid progenitors maintain an optimal balance between genetic diversity and allele fixation, setting the stage for interploidy crosses. In Chapter 4, we applied haplotype-based inbreeding tracking to quantify homozygosity levels in diploid material, offering a valuable tool for guiding parental selection.

Interploidy crosses: chromosome restitution and tetraploid selection

The final stage of Fix-Res breeding involves interploidy crosses, where diploid hybrid progenitors (HP) serve as male parents in crosses with tetraploid parents (TP). The presence of unreduced (2n) pollen ensures that the diploid chromosomal content is transmitted intact, resulting in tetraploid offspring that inherit all the genetic improvements from previous Fix-Res stages.

However, the tetraploid genetic contribution undergoes segregation, requiring selection to optimize its composition. PotatoMASH is used to genotype the tetraploid progeny, verifying the integrity of inherited alleles and tracking genetic variation introduced from the tetraploid parent. MAS and GS are applied to identify superior descendants, ensuring that only the most promising individuals are selected for commercialization as clonally propagated tetraploid varieties.

Genome-wide markers enable tracking of residual heterozygosity during inbreeding

Inbreeding plays a fundamental role in breeding programs by stabilizing traits and enabling allele fixation. However, not all genomic regions follow the expected pattern of progressive homozygosity across generations of selfing. Some loci remain persistently heterozygous, a phenomenon known as residual heterozygosity (RH), which is maintained due to selective pressures that favour heterozygous alleles (Marand et al. 2019). These regions often harbour alleles linked to beneficial agronomic traits such as fertility, vigour, and disease resistance, which may confer a disadvantage when fully homozygous (Marand et al. 2019; Peterson et al. 2016; Phumichai and Hosaka 2006). The persistence of RH presents a challenge in breeding programs, where achieving a balance between fixing beneficial alleles and maintaining genetic diversity is a critical objective.

Regions Resistant to Inbreeding

To investigate the persistence of RH across generations, we applied haplotype-based genotyping to assess homozygosity and heterozygosity frequencies at different stages of inbreeding in diploid clones from the Wageningen University & Research, Plant Breeding diploid breeding program. Our analysis confirmed the expected trend of increasing homozygosity across generations, but it also revealed a pattern of non-random

heterozygosity retention at specific loci, indicating that certain genomic regions are more resistant to fixation than others. These findings suggest that while inbreeding drives overall homozygosity, some loci remain selectively heterozygous due to functional constraints or adaptive advantages.

One of the most striking findings was the persistence of heterozygosity at locus C3_8 on chromosome 3, which remained heterozygous in 96% of the clones in this study, throughout the broader diploid breeding germplasm. While this locus has not been previously linked to a specific trait, chromosome 3 is known to harbour major QTL for yellow flesh colour (Bonierbale et al. 1988b) and has been reported as a hotspot for total protein content (Acharjee et al. 2018). In our GWAS study (Chapter 3), we identified several QTL related to yield components on chromosome 3, including a new QTL associated with low yield and low tuber number, which co-localized with a QTL for storage stability at cold temperatures. This storage QTL was associated with dark crisp colour, suggesting that genetic factors influencing both yield and post-harvest quality traits are linked to chromosome 3.

Similarly, locus C12_10, which remained heterozygous in 94% of individuals, is located near genes and QTL linked to various yield and reproductive traits on chromosome 12. These include QTL for yield (McCord et al. 2011) (Chapter 3), canopy vigour (Chapter 3), fruit set (Peterson et al. 2016), self-compatibility (Clot et al. 2020), and unreduced pollen production (Clot et al. 2024). The high retention of heterozygosity at these loci suggests that they may harbour deleterious alleles that are masked by heterozygosity, or alternatively, that heterozygosity confers an advantage in specific fitness-related traits. This provides further evidence that Fix-Res breeding must account for the selective pressures maintaining RH while optimizing allele inheritance for genetic gain.

Another notable observation was the behaviour of chromosome 5 in experimental progenies. Despite multiple generations of selfing, chromosome 5 remained fully heterozygous in the parents of each progeny. While some loci along chromosome 5 did reach homozygosity in some of the offspring, breeder inadvertently selected heterozygous individuals as parents for the next generation, suggesting an unconscious preference for plants that maintained heterozygosity at key loci. No individual in the study population reached full homozygosity across all loci on chromosome 5, indicating that this chromosome harbours genes under strong selective pressure. These findings align with previous studies identifying chromosome 5 as a key region for agronomic traits such as maturity (Kloosterman et al. 2013), canopy growth (Chapter 3), unreduced gamete production (Clot et al. 2024), tuber number, and overall yield (Marand et al. 2019). The continued selection of heterozygous individuals further reinforces the idea that RH is not merely a by-product of incomplete inbreeding but rather a functional component of plant fitness and selection dynamics.

Tracking alleles during inbreeding

A key finding of this study was that haplotype-based genotyping provides a more accurate measure of homozygosity than SNP-based methods. SNP-based homozygosity estimates tend to overestimate the extent of fixation because they fail to account for recombination

events and multi-allelic variation within genomic segments. In contrast, haplotag-based analysis provides a more precise view of genetic variation, allowing for the detection of regions where heterozygosity persists beyond Mendelian expectations.

When homozygosity levels were measured across different inbreeding stages broader diploid breeding germplasm, SNP-based analysis estimated an average homozygosity of 75.7% in parental lines, increasing to 83.1% in S1, 87.8% in S2, and 90.4% in S3 generations. However, when assessed using haplotags, homozygosity levels were significantly lower, with parental lines averaging 41.8% homozygosity, increasing to 59.1% in S1, 71.8% in S2, and 77.2% in S3. The discrepancy between these estimates indicates that SNP-based approaches may misrepresent inbreeding progression due to their reliance on single nucleotide changes rather than full haplotype structures. Unlike SNP-based methods, haplotag analysis does not require additional parental information to achieve accurate homozygosity measurements, further demonstrating its advantage in inbreeding studies.

The reason for these discrepancies lies in how each method processes genomic data. SNP-based homozygosity estimates often misinterpret IBS (identity-by-state) as IBD (identity-by-descent), leading to false assumptions of allele fixation. In contrast, haplotags reconstruct phased haplotypes from sequencing reads, distinguishing between true IBD and IBS events. This distinction is particularly valuable in highly heterozygous genomes such as potato, where the ability to accurately track recombination and allele segregation is useful for effective breeding strategies.

Implications for Fix-Res breeding

While RH presents a challenge in diploid F_1 hybrid breeding, Fix-Res leverages this phenomenon by incorporating RH into the selection process rather than eliminating it. Unlike conventional inbred-line breeding, Fix-Res allows for the strategic retention of heterozygosity at loci where it is beneficial, thereby mitigating inbreeding depression while still enabling genetic gain. This approach provides several advantages. First, it preserves adaptive alleles that contribute to fitness, yield, and reproductive success, ensuring that these important loci remain heterozygous rather than being lost due to selection pressure for homozygosity. Second, it helps maintain genetic stability by minimizing the risk of inbreeding depression, particularly in cases where homozygous alleles may have negative fitness consequences. Finally, it optimizes allele inheritance by selectively fixing beneficial alleles while avoiding the unintended fixation of deleterious alleles, ensuring that genetic gains are maximized without compromising adaptability.

By integrating haplotype-based inbreeding tracking, breeders can make data-driven decisions to selectively increase homozygosity at targeted loci while maintaining heterozygous regions essential for plant fitness. This sets Fix-Res apart from conventional inbred-line breeding, providing a strategic balance between fixation and genetic diversity, ultimately optimizing both genetic gain and population stability.

Future perspectives

In this thesis, I demonstrated the application of targeted genotyping via pooled multiple amplicon sequencing and backed-read haplotyping in potato breeding, showcasing its potential for cost-effective genotyping and haplotype-based analysis in GWAS, GS and Tracking alleles during Inbreeding. The findings presented here provide a foundation for integrating haplotype-based genomic tools into breeding pipelines, but further advancements are needed to refine and expand their application in Fix-Res and broader breeding programs. The following areas present key directions for the next phase of research and implementation.

Developing diagnostic markers and enhancing Genomic Selection

In this thesis, 37 QTL were identified for key agronomic and quality traits, providing a valuable resource for MAS and GS. However, before these markers can be effectively implemented in breeding programs, they must be validated across genetically diverse populations to ensure their reliability and consistency in selection. QTL validation is essential to confirm their predictive value across different genetic backgrounds.

As discussed in Chapters 3 and 5, while haplotypes provide greater genetic resolution compared to SNP-based approaches, their rarity can also reduce the amount of available information which can limit their statistical power in genetic analysis. To improve the robustness of haplotype-based selection, larger populations are needed to capture a wider range of haplotypic diversity and recombination patterns. Increasing population size and genetic diversity will also enhance the statistical power of QTL detection, ensuring that the markers identified in this study remain informative across different breeding panels.

Beyond QTL validation, expanding the training population for GS is another key priority. The accuracy of GEBVs depends on the diversity and representativeness of the training population, as well as the quality and breadth of phenotypic data. Future efforts should focus on incorporating a broader genetic pool and expanding multi-environment phenotypic evaluations to refine GP models. By strengthening both MAS and GS through larger and more diverse datasets, these markers can become powerful tools for accelerating genetic gain and improving selection efficiency in Fix-Res breeding.

Refining prediction strategies for Genomic Selection

In Chapter 5, we observed that traits controlled by major-effect QTL, such as Tuber Shape and Maturity, exhibited comparable prediction accuracies to polygenic traits like Yield, despite the intuitive expectation that strong marker-trait associations should lead to higher predictive accuracy. This finding suggests that current GP models may not be fully optimized for traits with major-effect loci, highlighting the need for alternative strategies to improve prediction efficiency.

Another key consideration is the role of rare alleles in GP models, particularly in genomic relationship matrix (GRM)-based approaches, which typically place more weight on common

variants. While rare alleles are often excluded due to their low frequency, they may still contribute significantly to trait variation, especially if they are linked to critical agronomic traits such as disease resistance. Given that haplotags identified in this study were often rare, improving the incorporation of rare alleles into prediction models is important for ensuring effective selection.

Future work should focus on testing alternative prediction models that account for both major-effect QTL and rare alleles. One potential strategy is placing greater weight on markers identified through GWAS or rare but favourable alleles in GEBV estimation, as suggested in previous studies (Byrne et al. 2020; Goddard 2009; Jannink 2010; Liu et al. 2014). Implementing such weighted approaches could enhance long-term genetic gain, particularly in breeding programs like Fix-Res, where maintaining adaptive genetic variation while optimizing selection efficiency is beneficial.

PotatoMASH Version 2.0

The current platform covers 339 loci spaced at 1 Mb intervals across the euchromatic portion of the genome, providing reasonable genome-wide coverage while being specifically designed as a low-cost genotyping solution. In Chapter 2, we demonstrated the adaptability of PotatoMASH, showing that its core marker set can be expanded to include additional diagnostic markers linked to important traits. Similar to the recent update of the PotatoDArTag V2 platform, an expanded version of PotatoMASH should facilitate the integration of newly discovered loci relevant to breeding programs.

Loci targeting resistance traits for MAS should be prioritized in the next version, particularly those associated with key disease resistance genes. In addition, incorporating the *Sli* locus on chromosome 12 would allow for MAS-based selection for self-compatibility, which is key for breeding at the diploid level. Expanding the marker set to include these critical loci would improve the platform's utility in both diploid and tetraploid breeding systems.

Beyond adding specific trait-linked markers, increasing the number of targeted loci by reducing the marker spacing from 1 Mb to 500 kb would significantly enhance the platform's resolution. A denser distribution of markers would improve the ability to detect marker-trait associations in GWAS and refine GP models. By capturing more recombination events and identifying additional QTL, a higher-density version of PotatoMASH would provide breeders with a more powerful tool for accelerating genetic gain.

Collaborating with a sequencing provider for automation and commercial implementation

Library preparation in the PotatoMASH workflow is a labour-intensive process involving multiple manual steps, including PCR reactions, barcode assignments, purification, and normalization, making it time-consuming and prone to errors. Automating these steps would significantly reduce hands-on time, minimize pipetting errors, and improve reproducibility, particularly by integrating multiplexed PCR reactions with automated normalization and pooling. The same applies to the bioinformatics pipeline, where handling sequenced reads

to call both SNP and haplotag marker sets requires high-performance computing (HPC) resources and bioinformatics expertise, whereas the final output can be easily processed in R on local computers. Additionally, some breeding programs, research facilities, and small commercial breeding companies lack the necessary in-house molecular or bioinformatics infrastructure, making them reliant on external genotyping services and manual library construction and data processing remain bottlenecks in terms of time and cost. To enhance scalability and accessibility, partnering with sequencing service providers or developing a more automation-friendly protocol would be beneficial. A particularly effective approach would be to establish PotatoMASH as a commercial service in collaboration with a genotyping provider like LGC Genomics (www.biosearchtech.com), which provide other Next-Generation Sequencing (NGS) services and has the molecular, sequencing, and bioinformatics capabilities to streamline the entire process, making it more efficient and accessible for routine breeding applications.

Developing bioinformatics tools for multi-allelic haplotypes

One of the primary advantages of haplotags identified through PotatoMASH is their ability to capture multi-allelic variation at a given locus, offering a more comprehensive representation of genetic diversity. However, in this study, haplotags were treated as pseudo-SNPs, meaning they were converted into binary or dosage-based formats rather than being analysed as true multi-allelic markers. This simplification was necessary to integrate PotatoMASH data into widely used genomic analysis tools such as GWASpoly (Rosyara et al. 2016) for GWAS and rrBLUP (Endelman 2011) or BGLR (Pérez and de los Campos 2014) for the development of GP models, all of which primarily support bi-allelic SNP-based models. While this approach proved effective in identifying QTL in Chapters 2 and 3 and achieving high GP accuracy in Chapter 5, it may have resulted in a loss of genetic information relevant for detecting complex trait associations.

Future research should focus on developing computational pipelines that can integrate haplotags as true multi-allelic markers, which could significantly enhance genetic analysis and improve selection efficiency. Understanding the nature of allelic interactions within and between loci would provide deeper insights into the genetic architecture of complex traits. Recently, Thérèse Navarro et al. (2022) developed mpQTL, a software for QTL analysis at any ploidy level under both bi-allelic and multi-allelic models, specifically designed for multi-parental populations. Their approach was demonstrated using simulated data of short-range haplotypes in autotetraploid multi-parental populations. Combining such methodologies with real-world data from genetically diverse panels like the current study could provide a clearer picture of the genetic control of traits in highly heterozygous systems. From both a genetic and practical breeding perspective, advancing these analytical frameworks would improve the precision of marker-assisted selection and accelerate genetic gains in potato breeding programs.

Concluding remarks:

This thesis demonstrates how a single genotyping platform can support key breeding activities in potato by combining cost-effective targeted amplicon sequencing with read-backed haplotyping. By facilitating GWAS, MAS, genomic prediction, and the tracking of inbreeding, this approach helps breeders make more informed decisions throughout the breeding process. Haplotags complement SNPs, improving QTL resolution and the tracking of inbreeding in highly heterozygous material. Genomic prediction using our low-density targeted genotyping assay achieves accuracies comparable to high-density marker data and perform well across a suite of complex traits. Although breeding by design remains a long-term goal in a complex crop like potato, this thesis contributes concrete steps toward that vision. The next steps will involve scaling up this approach to more diverse populations and expanding its use across breeding programs. Furthermore, implementing breeding strategies informed by genomic data will accelerate crop improvement and support the development of the next generation of potato varieties.

References

Aalborg T, Nielsen KL (2024) To be or not to be tetraploid—the impact of marker ploidy on genomic prediction and GWAS of potato. Frontiers in Plant Science 15:1386837. doi:10.3389/fpls.2024.1386837

Aalborg T, Sverrisdóttir E, Kristensen HT, Nielsen KL (2024) The effect of marker types and density on genomic prediction and GWAS of key performance traits in tetraploid potato. Frontiers in Plant Science 15:1340189. doi:10.3389/fpls.2024.1340189

Acharjee A, Chibon P-Y, Kloosterman B, America T, Renaut J, Maliepaard C, Visser RGF (2018) Genetical genomics of quality related traits in potato tubers using proteomics. BMC Plant Biology 18:20. doi:10.1186/s12870-018-1229-1

Adams J, de Vries M, van Eeuwijk F (2023) Efficient Genomic Prediction of Yield and Dry Matter in Hybrid Potato. Plants 12:2617. doi:10.3390/plants12142617

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403-410. doi:10.1016/S0022-2836(05)80360-2

Angelin-Bonnet O, Thomson S, Vignes M, Biggs PJ, Monaghan K, Bloomer R, Wright K, Baldwin S (2023) Investigating the genetic components of tuber bruising in a breeding population of tetraploid potatoes. BMC Plant Biology 23 (1):238. doi:10.1186/s12870-023-04255-2

Araujo AC, Carneiro PL, Oliveira HR, Lewis RM, Brito LF (2023) SNP-and haplotype-based single-step genomic predictions for body weight, wool, and reproductive traits in North American Rambouillet sheep. Journal of Animal Breeding and Genetics 140:216-234. doi:10.1111/jbg.12748

Asano K, Endelman JB (2024) Development of KASP Markers for the Potato Virus Y Resistance Gene Ry chc Using Whole-Genome Resequencing Data. American Journal of Potato Research 101 (2):114-121

Baird E, Cooper-Bland S, Waugh R, DeMaine M, Powell W (1992) Molecular characterisation of inter and intra-specific somatic hybrids of potato using randomly amplified polymorphic DNA (RAPD) markers. Molecular and General Genetics MGG 233 (3):469-475. doi:10.1007/BF00265445

Baldwin S, Dodds K, Auvray B, Genet R, Macknight R, Jacobs J (2011) Association mapping of cold-induced sweetening in potato using historical phenotypic data. Annals of Applied Biology 158 (3):248-256

Ballesta P, Maldonado C, Pérez-Rodríguez P, Mora F (2019) SNP and Haplotype-Based Genomic Selection of Quantitative Traits in Eucalyptus globulus. Plants 8:331. doi:10.3390/plants8090331

Bao Z, Li C, Li G, Wang P, Peng Z, Cheng L, Li H, Zhang Z, Li Y, Huang W, Ye M, Dong D, Cheng Z, VanderZaag P, Jacobsen E, Bachem CWB, Dong S, Zhang C, Huang S, Zhou Q (2022) Genome architecture and tetrasomic inheritance of autotetraploid potato. Molecular Plant 15:1211-1226. doi:10.1016/j.molp.2022.06.009

Baral K, Coulman B, Biligetu B, Fu Y-B (2020) Advancing crested wheatgrass [Agropyron cristatum (L.) Gaertn.] breeding through genotyping-by-sequencing and genomic selection. Plos one 15:e0239609. doi:10.1371/journal.pone.0239609

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265. doi:10.1093/bioinformatics/bth457

Bastien M, Boudhrioua C, Fortin G, Belzile F (2018) Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. Genome 61 (6):449-456

Bates D, Martin M, Ben B, Steve W (2015) Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67:1 - 48. doi:10.18637/jss.v067.i01

Batista LG, Mello VH, Souza AP, Margarido GRA (2022) Genomic prediction with allele dosage information in highly polyploid species. Theoretical and Applied Genetics 135:723-739. doi:10.1007/s00122-021-03994-w

Bernardo A, St. Amand P, Le HQ, Su Z, Bai G (2020) Multiplex restriction amplicon sequencing: a novel next-generation sequencing-based marker platform for high-throughput genotyping. Plant Biotechnology Journal 18 (1):254-265. doi:https://doi.org/10.1111/pbi.13192

Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK (2021) Features and applications of haplotypes in crop breeding. Communications Biology 4 (1):1266. doi:10.1038/s42003-021-02782-y

Bonierbale MW, Plaisted RL, Tanksley S (1988a) RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics 120:1095-1103. doi:https://doi.org/10.1093/genetics/120.4.1095

Bonierbale MW, Plaisted RL, Tanksley SD (1988b) RFLP Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato. Genetics 120 (4):1095-1103. doi:10.1093/genetics/120.4.1095

Bourke PM, van Geest G, Voorrips RE, Jansen J, Kranenburg T, Shahin A, Visser RGF, Arens P, Smulders MJM, Maliepaard C (2018) polymapR—linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. Bioinformatics 34:3496-3502. doi:10.1093/bioinformatics/bty371

Bradshaw JE (2007) The Canon of Potato Science: 4. Tetrasomic Inheritance. Potato Research 50 (3):219-222. doi:10.1007/s11540-008-9041-1

Bradshaw JE (2017) Review and Analysis of Limitations in Ways to Improve Conventional Potato Breeding. Potato Research 60 (2):171-193. doi:10.1007/s11540-017-9346-z

Bradshaw JE (2022) Breeding diploid F1 hybrid potatoes for propagation from botanical seed (TPS): comparisons with theory and other crops. Plants 11:1121. doi:https://doi.org/10.3390/plants11091121

Bradshaw JE, Hackett CA, Pande B, Waugh R, Bryan GJ (2008) QTL mapping of yield, agronomic and quality traits in tetraploid potato (Solanum tuberosum subsp. tuberosum). Theoretical and Applied Genetics 116:193-211. doi:https://10.1007/s00122-007-0659-1

Braun SR, Endelman JB, Haynes KG, Jansky SH (2017) Quantitative Trait Loci for Resistance to Common Scab and Cold-Induced Sweetening in Diploid Potato. The Plant Genome 10 (3):plantgenome2016.2010.0110. doi:https://doi.org/10.3835/plantgenome2016.10.0110

Brinton J, Ramirez-Gonzalez RH, Simmonds J, Wingen L, Orford S, Griffiths S, Project WG, Haberer G, Spannagl M, Walkowiak S (2020) A haplotype-led approach to increase the precision of wheat breeding. Communications Biology 3 (1):712

Brown C, Kim T, Ganga Z, Haynes K, De Jong D, Jahn M, Paran I, De Jong W (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. doi:https://doi.org/10.1007/BF02872013

Brown SS, Chen Y-W, Wang M, Clipson A, Ochoa E, Du M-Q (2017) PrimerPooler: automated primer pooling to prepare library for targeted sequencing. Biology Methods and Protocols 2:bpx006. doi:10.1093/biomethods/bpx006

Browning BL, Browning SR (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics 84:210-223. doi:10.1016/j.ajhg.2009.01.005

Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics 81:1084-1097. doi:10.1086/521987

Byrne S, Czaban A, Studer B, Panitz F, Bendixen C, Asp T (2013) Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. PloS one 8:e57438. doi:10.1371/journal.pone.0057438

Byrne S, Meade F, Mesiti F, Griffin D, Kennedy C, Milbourne D (2020) Genome-Wide Association and Genomic Prediction for Fry Color in Potato. Agronomy 10:90. doi:10.3390/agronomy10010090

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178:553-561. doi:10.1534/genetics.107.080838

Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular ecology resources 15:855-867. doi:10.1111/1755-0998.12357

Canales FJ, Montilla-Bascón G, Bekele WA, Howarth CJ, Langdon T, Rispail N, Tinker NA, Prats E (2021) Population genomics of Mediterranean oat (A. sativa) reveals high genetic diversity and three loci for heading date. Theoretical and Applied Genetics 134:2063-2077. doi:10.1007/s00122-021-03805-2

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nature Reviews Genetics 10:783-796. doi:10.1038/nrg2664

Chase SS (1963) Analytic breeding in *Solanum tuberosum* L.: a scheme utilizing parthenotes and other diploid stocks. Canadian Journal of Genetics and Cytology 5:359-363. doi:https://doi.org/10.1139/g63-049

Clot CR (2023) Natural variation in potato sexual reproduction facilitates breeding. Dissertation, Wageningen University,

Clot CR, Klein D, Koopman J, Schuit C, Engelen CJM, Hutten RCB, Brouwer M, Visser RGF, Juranić M, van Eck HJ (2023) Desynapsis in potato is caused by *StMSH4* mutant alleles and leads to either highly uniform unreduced pollen or sterility. bioRxiv:2023.2002.2023.529759. doi:https://doi.org/10.1101/2023.02.23.529759

Clot CR, Polzer C, Prodhomme C, Schuit C, Engelen CJM, Hutten RCB, van Eck HJ (2020) The origin and widespread occurrence of *Sli*-based self-compatibility in potato. Theoretical and Applied Genetics 133:2713-2728. doi:https://10.1007/s00122-020-03627-8

Clot CR, Vexler L, Leyva-Pérez MdlO, Bourke PM, Engelen CJM, Hutten RCB, van de Belt J, Wijnker E, Milbourne D, Visser RGF, Juranić M, van Eck HJ (2024) Identification of two mutant JASON-RELATED genes associated with unreduced pollen production in potato. Theoretical and Applied Genetics 137:79. doi:10.1007/s00122-024-04563-7

Consortium TPGS (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189-195. doi:10.1038/nature10158

Cuyabano BCD, Su G, Lund MS (2014) Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. BMC Genomics 15:1171. doi:10.1186/1471-2164-15-1171

D'hoop BB, Keizer PLC, Paulo MJ, Visser RGF, van Eeuwijk FA, van Eck HJ (2014) Identification of agronomically important QTL in tetraploid potato cultivars using a marker–trait association analysis. Theoretical and Applied Genetics 127:731-748. doi:https://10.1007/s00122-013-2254-y

D'hoop BB, Paulo MJ, Mank RA, van Eck HJ, van Eeuwijk FA (2008) Association mapping of quality traits in potato (*Solanum tuberosum* L.). Euphytica 161:47-60. doi:https://10.1007/s10681-007-9565-5

da Silva Pereira G, Mollinari M, Qu X, Thill C, Zeng Z-B, Haynes K, Yencho GC (2021a) Quantitative Trait Locus Mapping for Common Scab Resistance in a Tetraploid Potato Full-Sib Population. Plant Disease 105 (10):3048-3054. doi:10.1094/PDIS-10-20-2270-RE

da Silva Pereira G, Mollinari M, Schumann MJ, Clough ME, Zeng Z-B, Yencho GC (2021b) The recombination landscape and multiple QTL mapping in a Solanum tuberosum cv. 'Atlantic'-derived F1 population. Heredity 126 (5):817-830. doi:10.1038/s41437-021-00416-x

De Jong H, Rowe P (1971a) Inbreeding in cultivated diploid potatoes. Potato Research 14:74-83

De Jong H, Rowe PR (1971b) Inbreeding in cultivated diploid potatoes. Potato Research 14 (2):74-83. doi:10.1007/BF02355931

Demeke T, Lynch DR, Kawchuk LM, Kozub GC, Armstrong JD (1996) Genetic diversity of potato determined by random amplified polymorphic DNA analysis. Plant Cell Reports 15 (9):662-667. doi:10.1007/BF00231920

Develtere W, Waegneer E, Debray K, De Saeger J, Van Glabeke S, Maere S, Ruttink T, Jacobs TB (2023) *SMAP design*: a multiplex PCR amplicon and gRNA design tool to screen for natural and CRISPR-induced genetic variation. Nucleic acids research 51:e37. doi:10.1093/nar/gkad036

Díaz P, Sarmiento F, Mathew B, Ballvora A, Mosquera Vásquez T (2021) Genomic regions associated with physiological, biochemical and yield-related responses under water deficit in diploid potato at the tuber initiation stage revealed by GWAS. PLOS ONE 16 (11):e0259690. doi:10.1371/journal.pone.0259690

Difabachew YF, Frisch M, Langstroff AL, Stahl A, Wittkop B, Snowdon RJ, Koch M, Kirchhoff M, Cselényi L, Wolf M, Förster J, Weber S, Okoye UJ, Zenke-Philippi C (2023) Genomic prediction with haplotype blocks in wheat. Frontiers in Plant Science 14:1168547. doi:10.3389/fpls.2023.1168547

Dodds KG, McEwan JC, Brauning R, Anderson RM, van Stijn TC, Kristjánsson T, Clarke SM (2015) Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics 16:1047. doi:10.1186/s12864-015-2252-3

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one 6:e19379. doi:10.1371/journal.pone.0019379

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. The plant genome 4. doi:10.3835/plantgenome2011.08.0024

Endelman JB, Kante M, Lindqvist-Kreuze H, Kilian A, Shannon LM, Caraza-Harter MV, Vaillancourt B, Mailloux K, Hamilton JP, Buell CR (2024) Targeted genotyping-by-sequencing of potato and data analysis with R/polyBreedR. The Plant Genome 17:e20484. doi:10.1002/tpg2.20484

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Molecular ecology 14:2611-2620. doi:https://10.1111/j.1365-294X.2005.02553.x

Ewing E, Struik P (1992) Tuber formation in potato: induction, initiation, and growth. Horticultural reviews 14:89-198. doi:10.1002/9780470650523.ch3

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

Fox J, Weisberg S (2019) An R Companion to Applied Regression. Sage, Thousand Oaks CA

Fradgley N, Gardner KA, Cockram J, Elderfield J, Hickey JM, Howell P, Jackson R, Mackay IJ (2019) A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. PLoS Biology 17 (2):e3000071

Garrison E, Marth G (2016) Haplotype-based variant detection from short-read sequencing. arXiv 1207. doi:10.48550/arXiv.1207.3907

Gebhardt C (2013) Bridging the gap between genome analysis and precision breeding in potato. Trends in Genetics 29 (4):248-256

Gebhardt C, Mugniery D, Ritter E, Salamini F, Bonnel E (1993) Identification of RFLP markers closely linked to the H1 gene conferring resistance to Globodera rostochiensis in potato. Theoretical and Applied Genetics 85 (5):541-544. doi:10.1007/BF00220911

Gebhardt C, Ritter E, Barone A, Debener T, Walkemeier B, Schachtschabel U, Kaufmann H, Thompson RD, Bonierbale MW, Ganal MW, Tanksley SD, Salamini F (1991) RFLP maps of potato and their alignment with the homoeologous tomato genome. Theoretical and Applied Genetics 83 (1):49-57. doi:10.1007/BF00229225

Ghislain M, Douches DS (2020) The Genes and Genomes of the Potato. In: Campos H, Ortiz O (eds) The Potato Crop: Its Agricultural, Nutritional and Social Contribution to Humankind. Springer International Publishing, Cham, pp 139-162. doi:10.1007/978-3-030-28683-5_5

Ghislain M, Spooner DM, Rodríguez F, Villamón F, Núñez J, Vásquez C, Waugh R, Bonierbale M (2004) Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. Theoretical and Applied Genetics 108 (5):881-890. doi:10.1007/s00122-003-1494-7

Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive Genetic Variability and the Bayesian Alphabet. Genetics 183:347-363. doi:10.1534/genetics.109.103952

Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257. doi:10.1007/s10709-008-9308-0

Goddard M, Hayes B (2007) Genomic selection. Journal of Animal breeding and Genetics 124:323-330. doi:10.1111/j.1439-0388.2007.00702.x.

Gordon AH, G. (2010) Fastx-Toolkit. FASTQ/A Short-Reads Pre-Processing Tools. Available online: http://hannonlab.cshl.edu/fastx_toolkit (accessed on 26 August 2022).

Grech-Baran M, Witek K, Szajko K, Witek AI, Morgiewicz K, Wasilewicz-Flis I, Jakuczun H, Marczewski W, Jones JDG, Hennig J (2020) Extreme resistance to Potato virus Y in potato carrying the Ry gene is mediated by a TIR-NLR immune receptor. Plant Biotechnology Journal 18:655-667. doi:10.1111/pbi.13230

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. doi:10.1186/1471-2105-12-186

Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, Van Deynze A, De Jong WS, Douches DS, Buell CR (2011) Single nucleotide polymorphism discovery in elite north american potato germplasm. BMC Genomics 12:302. doi:10.1186/1471-2164-12-302

Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B, Yang X, Zeng Z, Douches DS, Jiang J, Veilleux RE, Buell CR (2016) Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum The Plant Cell 28:388-405. doi:10.1105/tpc.15.00538

Hardigan MA, Feldmann MJ, Carling J, Zhu A, Kilian A, Famula RA, Cole GS, Knapp SJ (2023) A medium-density genotyping platform for cultivated strawberry using DArTag technology. The Plant Genome 16 (4):e20399. doi:https://doi.org/10.1002/tpg2.20399

Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53:876-883. doi:10.1139/G10-076

Hayes B, Goddard ME (2001) The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution 33:209. doi:10.1186/1297-9686-33-3-209

He C, Holme J, Anthony J (2014) SNP genotyping: the KASP assay. In: Delphine Fleury RW (ed) Crop breeding, vol 1145. Methods in Molecular Biology, 1 edn. Humana Press, New York, NY, pp 75-86. doi:10.1007/978-1-4939-0446-4_7

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Science 49 (1):1-12. doi:10.2135/cropsci2008.08.0512

Hess M, Druet T, Hess A, Garrick D (2017) Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genetics Selection Evolution 49:54. doi:10.1186/s12711-017-0329-y

Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS genetics 4:e1000008. doi:10.1371/journal.pgen.1000008

Hosaka K, Hanneman RE (1998a) Genetics of self-compatibility in a self-incompatible wild diploid potato species *Solanum chacoense*. 1. Detection of an *S* locus inhibitor (*Sli*) gene. Euphytica 99:191-197. doi:https://doi.org/10.1023/A:1018353613431

Hosaka K, Hanneman RE (1998b) 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. doi:https://10.1023/A:1018380725160

Hosaka K, Sanetomo R (2020) Creation of a highly homozygous diploid potato using the S locus inhibitor (Sli) gene. Euphytica 216:169. doi:10.1007/s10681-020-02699-3

Hougas RW, Peloquin SJ (1958) The potential of potato haploids in breeding and genetic research. American Potato Journal 35 (10):701-707. doi:10.1007/BF02855564

Hu Q, Tang C, Zhou X, Yang X, Luo Z, Wang L, Yang M, Li D, Li L (2023) Potatoes dormancy release and sprouting commencement: A review on current and future prospects. Food Frontiers 4:1001-1018. doi:10.1002/fft2.228

Hutten R, Schippers M, Hermsen JT, Ramanna M (1994) Comparative performance of FDR and SDR progenies from reciprocal 4x-2x crosses in potato. Theoretical and Applied Genetics 89:545-550

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Research 29 (4):e25-e25. doi:10.1093/nar/29.4.e25

Jackson MT (1987) Breeding strategies for true potato seed. The production of new varieties: technological advances:248-261

Jannink J-L (2010) Dynamics of long-term genomic selection. Genetics Selection Evolution 42 (1):35. doi:10.1186/1297-9686-42-35

Jansky S, Chung Y, Kittipadukal P (2014) M6: A diploid potato inbred line for use in breeding and genetics research. Journal of Plant Registrations 8 (2):195-199

Jansky SH, Charkowski AO, Douches DS, Gusmini G, Richael C, Bethke PC, Spooner DM, Novy RG, De Jong H, De Jong WS, Bamberg JB, Thompson AL, Bizimungu B, Holm DG, Brown CR, Haynes KG, Sathuvalli VR, Veilleux RE, Miller Jr JC, Bradeen JM, Jiang J (2016) Reinventing Potato as a Diploid Inbred Line–Based Crop. Crop Science 56 (4):1412-1422. doi:https://doi.org/10.2135/cropsci2015.12.0740

Jiang Y, Schmidt RH, Reif JC (2018) Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. G3 Genes|Genomes|Genetics 8:1687-1699. doi:10.1534/g3.117.300548

Jupe F, Witek K, Verweij W, Śliwka J, Pritchard L, Etherington GJ, Maclean D, Cock PJ, Leggett RM, Bryan GJ, Cardle L, Hein I, Jones JDG (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. The Plant Journal 76:530-544. doi:10.1111/tpj.12307

Kacheyo OC, van Dijk LCM, de Vries ME, Struik PC (2021) Augmented descriptions of growth and development stages of potato (*Solanum tuberosum* L.) grown from different types of planting material. Annals of Applied Biology 178:549-566. doi:10.1111/aab.12661

Kaiser N, Jansky S, Coombs J, Collins P, Alsahlany M, Douches D (2021) Assessing the Contribution of Sli to Self-Compatibility in North American Diploid Potato Germplasm Using KASP™ Markers. American Journal of Potato Research 98. doi:10.1007/s12230-021-09821-8

Kaiser NR, Coombs JJ, Felcher KJ, Hammerschmidt R, Zuehlke ML, Buell CR, Douches DS (2020) Genome-Wide Association Analysis of Common Scab Resistance and Expression Profiling of Tubers in Response to Thaxtomin A Treatment Underscore the Complexity of Common Scab Resistance in Tetraploid Potato. American Journal of Potato Research 97 (5):513-522. doi:10.1007/s12230-020-09800-5

Kang J (2023) Making Better Use of Genotyping-By-Sequencing (GBS) Data to Improve Perennial Ryegrass Breeding. Dissertation, University of Otago,

Kao TH, McCubbin AG (1996) How flowering plants discriminate between self and non-self pollen to prevent inbreeding. Proceedings of the National Academy of Sciences 93:12059-12065. doi:https://10.1073/pnas.93.22.12059

Kidane-Mariam H-M, Peloquin S (1975) Method of diplandroid formation and yield of progeny from reciprocal (4x-2x) crosses.

Klaassen MT, Willemsen JH, Vos PG, Visser RG, van Eck HJ, Maliepaard C, Trindade LM (2019) Genome-wide association analysis in tetraploid potato reveals four QTLs for protein content. Molecular Breeding 39:1-12

Kloosterman B (2006) Transcriptomic analysis of potato tuber development and tuber quality traits using microarray technology. Dissertation, Wageningen University,

Kloosterman B, Abelenda JA, Gomez MdMC, Oortwijn M, de Boer JM, Kowitwanich K, Horvath BM, van Eck HJ, Smaczniak C, Prat S, Visser RGF, Bachem CWB (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495:246-250. doi:10.1038/nature11912

Kloosterman B, De Koeyer D, Griffiths R, Flinn B, Steuernagel B, Scholz U, Sonnewald S, Sonnewald U, Bryan GJ, Prat S, Bánfalvi Z, Hammond JP, Geigenberger P, Nielsen KL, Visser RGF, Bachem CWB (2008) Genes driving potato tuber initiation and growth: identification based on transcriptional changes using the POCI array. Functional & Integrative Genomics 8:329-340. doi:10.1007/s10142-008-0083-x

Kloosterman B, Oortwijn M, Uitdewilligen J, America T, de Vos R, Visser RG, Bachem CW (2010) From QTL to candidate gene: genetical genomics of simple and complex traits in potato using a pooling strategy. BMC genomics 11:1-16. doi:10.1186/1471-2164-11-158

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PloS one 6:e15925. doi:10.1371/journal.pone.0015925

Krantz FA (1924) Potato breeding methods. Minn Agric Exp Stn Tech Bull 25:1–32

Krantz FA (1946) Potato breeding methods 3: a suggested procedure for potato breeding. Minn Agric Exp Stn Tech Bull 173:1-26

Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, Keurentjes JJB, van Eeuwijk FA (2014) Marker-Based Estimation of Heritability in Immortal Populations. Genetics 199 (2):379-398. doi:https://doi.org/10.1534/genetics.114.167916

Kruijer W, White I (2023) heritability: Marker-Based Estimation of Heritability Using Individual Plant or Plot Data.

Ledesma-Ramírez L, Solís-Moya E, Iturriaga G, Sehgal D, Reyes-Valdes MH, Montero-Tavera V, Sansaloni CP, Burgueño J, Ortiz C, Aguirre-Mancilla CL, Ramírez-Pimentel JG, Vikram P, Singh S (2019) GWAS to Identify Genetic Loci for Resistance to Yellow Rust in Wheat Pre-Breeding Lines Derived From Diverse Exotic Crosses. Frontiers in Plant Science 10. doi:10.3389/fpls.2019.01390

Leisner CP, Hamilton JP, Crisovan E, Manrique-Carpintero NC, Marand AP, Newton L, Pham GM, Jiang J, Douches DS, Jansky SH, Buell CR (2018) Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species Solanum chacoense, reveals residual heterozygosity. The Plant Journal 94 (3):562-570. doi:https://doi.org/10.1111/tpj.13857

Lenth R, Singmann H, Love J, Buerkner P, Herve M (2021) Emmeans: Estimated marginal means, aka least-squares means. R Package. https://github.com/rvlenth/emmeans.

Leonel M, Do Carmo EL, Fernandes AM, Soratto RP, Ebúrneo JAM, Garcia ÉL, Dos Santos TPR (2017) Chemical composition of potato tubers: the effect of cultivars and growth conditions. Journal of food science and technology 54:2372-2378. doi:10.1007/s13197-017-2677-6

Leyva-Pérez MdlO, Vexler L, Byrne S, Clot CR, Meade F, Griffin D, Ruttink T, Kang J, Milbourne D (2022) PotatoMASH - A Low Cost, Genome-Scanning Marker System for Use in Potato Genomics and Genetics Applications. Agronomy 12:2461. doi:10.3390/agronomy12102461

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-2993. doi:10.1093/bioinformatics/btr509

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. doi:10.48550/arXiv.1303.3997

Li H, Wang Z, Xu L, Li Q, Gao H, Ma H, Cai W, Chen Y, Gao X, Zhang L (2022) Genomic prediction of carcass traits using different haplotype block partitioning methods in beef cattle. Evolutionary Applications 15:2028-2042. doi:10.1111/eva.13491

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

Li X-Q, De Jong H, De Jong DM, De Jong WS (2005) Inheritance and genetic mapping of tuber eye depth in cultivated diploid potatoes. Theoretical and Applied Genetics 110:1068-1073. doi:https://10.1007/s00122-005-1927-6

Liaw A, Wiener M (2002) Classification and regression by randomForest. R news 2:18-22

Lindhout P, Meijer D, Schotte T, Hutten RCB, Visser RGF, van Eck HJ (2011) Towards F1 Hybrid Seed Potato Breeding. Potato Research 54:301-312. doi:https://10.1007/s11540-011-9196-z

Lindqvist-Kreuze H, Gastelo M, Perez W, Forbes GA, de Koeyer D, Bonierbale M (2014) Phenotypic stability and genome-wide association study of late blight resistance in potato genotypes adapted to the tropical highlands. Phytopathology 104 (6):624-633

Liu H, Sørensen AC, Berg P Optimum contribution selection combined with weighting rare favourable alleles increases long-term genetic gain. In: 10th World Congress on Genetics Applied to Livestock Production (WCGALP), 2014.

Lorenzo CD, Debray K, Herwegh D, Develtere W, Impens L, Schaumont D, Vandeputte W, Aesaert S, Coussens G, De Boe Y (2023) BREEDIT: a multiplex genome editing strategy to improve complex quantitative traits in maize. The Plant Cell 35:218-238. doi:10.1093/plcell/koac243

Lu Y, Shah T, Hao Z, Taba S, Zhang S, Gao S, Liu J, Cao M, Wang J, Prakash AB (2011) Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. PloS one 6 (9):e24861

Lu Y, Xu J, Yuan Z, Hao Z, Xie C, Li X, Shah T, Lan H, Zhang S, Rong T, Xu Y (2012) Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. Molecular Breeding 30 (1):407-418. doi:10.1007/s11032-011-9631-5

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957-2963. doi:10.1093/bioinformatics/btr507

Malosetti M, Van der Linden C, Vosman B, Van Eeuwijk F (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175 (2):879-889

Manrique-Carpintero NC, Coombs JJ, Cui Y, Veilleux RE, Buell CR, Douches D (2015) Genetic Map and QTL Analysis of Agronomic Traits in a Diploid Potato Population using Single Nucleotide Polymorphism Markers. Crop Science 55:2566-2579. doi:https://doi.org/10.2135/cropsci2014.10.0745

Marand AP, Jansky SH, Gage JL, Hamernik AJ, de Leon N, Jiang J (2019) Residual Heterozygosity and Epistatic Interactions Underlie the Complex Genetic Architecture of Yield in Diploid Potato. Genetics 212 (1):317-332. doi:10.1534/genetics.119.302036

McCord PH, Sosinski BR, Haynes KG, Clough ME, Yencho GC (2011) Linkage Mapping and QTL Analysis of Agronomic Traits in Tetraploid Potato (*Solanum tuberosum subsp. tuberosum*). Crop Science 51:771-785. doi:https://doi.org/10.2135/cropsci2010.02.0108

Meade F, Byrne S, Griffin D, Kennedy C, Mesiti F, Milbourne D (2020a) Rapid Development of KASP Markers for Disease Resistance Genes Using Pooled Whole-Genome Resequencing. Potato Research 63:57-73. doi:10.1007/s11540-019-09428-x

Meade F, Byrne S, Griffin D, Kennedy C, Mesiti F, Milbourne D (2020b) Rapid Development of KASP Markers for Disease Resistance Genes Using Pooled Whole-Genome Resequencing. Potato Research 63 (1):57-73. doi:10.1007/s11540-019-09428-x

Meade F, Hutten R, Wagener S, Prigge V, Dalton E, Kirk HG, Griffin D, Milbourne D (2020c) Detection of novel QTLs for late blight resistance derived from the wild potato species Solanum microdontum and Solanum pampasense. Genes 11:732. doi:10.3390/genes11070732

Medina CA, Zhao D, Lin M, Sapkota M, Sandercock AM, Beil CT, Sheehan MJ, Irish BM, Yu L-X, Poudel H, Claessens A, Moore V, Crawford J, Hansen J, Viands D, Peel MD, Tilhou N, Riday H, Brummer EC, Xu Z (2025) Pre-breeding in alfalfa germplasm develops highly differentiated populations, as revealed by genome-wide microhaplotype markers. Scientific Reports 15 (1):1253. doi:10.1038/s41598-024-84262-x

Meher PK, Rustgi S, Kumar A (2022) Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. Heredity 128:519-530. doi:10.1038/s41437-022-00539-9

Mendiburu A, Peloquin S (1971) High-yielding tetraploids from 4x-2x and 2x-2x matings.

Mendiburu AO, Peloquin SJ (1977) The significance of 2N gametes in potato breeding. Theoretical and Applied Genetics 49 (2):53-61. doi:10.1007/BF00275164

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819-1829. doi:10.1093/genetics/157.4.1819

Meuwissen THE, Odegard J, Andersen-Ranberg I, Grindflek E (2014) On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. Genetics Selection Evolution 46:49. doi:10.1186/1297-9686-46-49

Meuwissen THE, Solberg TR, Shepherd R, Woolliams JA (2009) A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genetics Selection Evolution 41:2. doi:10.1186/1297-9686-41-2

Milbourne D, Meyer RC, Collins AJ, Ramsay LD, Gebhardt C, Waugh R (1998) Isolation, characterisation and mapping of simple sequence repeat loci in potato. Molecular and General Genetics MGG 259 (3):233-245. doi:10.1007/s004380050809

Mosquera T, Alvarez MF, Jiménez-Gómez JM, Muktar MS, Paulo MJ, Steinemann S, Li J, Draffehn A, Hofmann A, Lübeck J (2016) Targeted and untargeted approaches unravel novel candidate genes and diagnostic SNPs for quantitative resistance of the potato (*Solanum tuberosum* L.) to Phytophthora infestans causing the late blight disease. PLoS One 11:e0156254. doi:https://doi.org/10.1371/journal.pone.0156254

Navarro C, Abelenda JA, Cruz-Oró E, Cuéllar CA, Tamaki S, Silva J, Shimamoto K, Prat S (2011) Control of flowering and storage organ formation in potato by FLOWERING LOCUS T. Nature 478:119-122. doi:10.1038/nature10431

Olsder J, Hermsen JGT (1976) Genetics of self-compatibility in dihaploids of Solanum tuberosum L. I. Breeding behaviour of two self-compatible dihaploids. Euphytica 25:597-607. doi:10.1007/BF00041597

Ortiz R, Crossa J, Reslow F, Perez-Rodriguez P, Cuevas J (2022) Genome-Based Genotype × Environment Prediction Enhances Potato (Solanum tuberosum L.) Improvement Using Pseudo-Diploid and Polysomic Tetraploid Modeling. Frontiers in Plant Science 13:785196. doi:10.3389/fpls.2022.785196

Ortiz R, Mihovilovich E (2020) Genetics and Cytogenetics of the Potato. In: Campos H, Ortiz O (eds) The Potato Crop: Its Agricultural, Nutritional and Social Contribution to Humankind. Springer International Publishing, Cham, pp 219-247. doi:10.1007/978-3-030-28683-5_7

Ortiz R, Peloquin SJ, Freyre R, Iwanaga M (1991) Efficiency of potato breeding using FDR 2n gametes for multitrait selection and progeny testing. Theoretical and Applied Genetics 82:602-608

Ortiz R, Reslow F, Montesinos-López A, Huicho J, Pérez-Rodríguez P, Montesinos-López OA, Crossa J (2023) Partial least squares enhance multi-trait genomic prediction of potato cultivars in new environments. Scientific Reports 13:9947. doi:10.1038/s41598-023-37169-y

Pandey J, Scheuring DC, Koym JW, Endelman JB, Vales MI (2023) Genomic selection and genome-wide association studies in tetraploid chipping potatoes. The Plant Genome 16:e20297. doi:10.1002/tpg2.20297

Pandey J, Scheuring DC, Koym JW, Vales MI (2022) Genomic regions associated with tuber traits in tetraploid potatoes and identification of superior clones for breeding purposes. Frontiers in Plant Science 13:952263. doi:https://10.3389/fpls.2022.952263

Park T, Casella G (2008) The bayesian lasso. Journal of the American Statistical Association 103:681-686. doi:10.1198/016214508000000337

Parra-Galindo MA, Soto-Sedano JC, Mosquera-Vásquez T, Roda F (2021) Pathway-based analysis of anthocyanin diversity in diploid potato. PLOS ONE 16 (4):e0250861. doi:10.1371/journal.pone.0250861

Peleman JD, van der Voort JR (2003) Breeding by Design. Trends in Plant Science 8 (7):330-334. doi:https://doi.org/10.1016/S1360-1385(03)00134-1

Pérez P, de los Campos G (2014) Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics 198:483-495. doi:10.1534/genetics.114.164442

Peterson BA, Holt SH, Laimbeer FPE, Doulis AG, Coombs J, Douches DS, Hardigan MA, Buell CR, Veilleux RE (2016) Self-Fertility in a Cultivated Diploid Potato Population Examined with the Infinium 8303 Potato Single-Nucleotide Polymorphism Array. The Plant Genome 9 (3):plantgenome2016.2001.0003. doi:https://doi.org/10.3835/plantgenome2016.01.0003

Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, Ou S, Jiang J, Buell CR (2020) Construction of a chromosome-scale long-read reference genome assembly for potato. GigaScience 9:p.giaa100. doi:https://10.1093/gigascience/giaa100

Phumichai C, Hosaka K (2006) Cryptic improvement for fertility by continuous selfing of diploid potatoes using Sli gene. Euphytica 149 (1):251-258. doi:10.1007/s10681-005-9072-5

Phumichai C, Mori M, Kobayashi A, Kamijima O, Hosaka K (2005) Toward the development of highly homozygous diploid potato lines using the self-compatibility controlling Sli gene. Genome 48 (6):977-984. doi:10.1139/g05-066

Powell JE, Visscher PM, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11 (11):800-805. doi:10.1038/nrg2865

Preedy KF, Hackett CA (2016) A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. Theoretical and Applied Genetics 129:2117-2132. doi:10.1007/s00122-016-2761-8

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945-959. doi:https://10.1093/genetics/155.2.945

Prodhomme C, Vos PG, Paulo MJ, Tammes JE, Visser RG, Vossen JH, van Eck HJ (2020) Distribution of P1 (D1) wart disease resistance in potato germplasm and GWAS identification of haplotype-specific SNP markers. Theoretical and Applied Genetics 133:1859-1871. doi:10.1007/s00122-020-03559-3

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Bakker P, Daly M, Sham P (2007) Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American journal of human genetics 81:559-575. doi:https://10.1086/519795

Qu L, Huang X, Su X, Zhu G, Zheng L, Lin J, Wang J, Xue H (2024) Potato: from functional genomics to genetic improvement. Molecular Horticulture 4:34. doi:10.1186/s43897-024-00105-3

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics 197 (2):573-589. doi:https://10.1534/genetics.114.164350

Rak K, Bethke PC, Palta JP (2017) QTL mapping of potato chip color and tuber traits within an autotetraploid family. Molecular Breeding 37:15. doi:https://10.1007/s11032-017-0619-7

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nature biotechnology 29:24-26. doi:10.1038/nbt.1754

Rosyara UR, De Jong WS, Douches DS, Endelman JB (2016) Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. The Plant Genome 9:plantgenome2015.2008.0073. doi:10.3835/plantgenome2015.08.0073

Rouppe van der Voort J, van der Vossen E, Bakker E, Overmars H, van Zandvoort P, Hutten R, Klein Lankhorst R, Bakker J (2000) Two additive QTLs conferring broad-spectrum resistance in potato to Globodera pallida are localized on resistance gene clusters. Theoretical and Applied Genetics 101:1122-1130. doi:10.1007/s001220051588

Sallam AH, Conley E, Prakapenka D, Da Y, Anderson JA (2020) Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. G3 Genes|Genomes|Genetics 10:2265-2273. doi:10.1534/g3.120.401165

Schaumont D, Veeckman E, Van der Jeugt F, Haegeman A, van Glabeke S, Bawin Y, Lukasiewicz J, Blugeon S, Barre P, Leyva-Pérez MdlO, Byrne S, Dawyndt P, Ruttink T (2022) Stack Mapping Anchor Points (SMAP): a versatile suite of tools for read-backed haplotyping. bioRxiv:2022.2003.2010.483555. doi:10.1101/2022.03.10.483555

Schönhals E, Ortega F, Barandalla L, Aragones A, Ruiz de Galarreta J, Liao J-C, Sanetomo R, Walkemeier B, Tacke E, Ritter E (2016) Identification and reproducibility of diagnostic DNA markers for tuber starch and yield optimization in a novel association mapping population of potato (Solanum tuberosum L.). Theoretical and Applied Genetics 129:767-785

Schönhals EM, Ding J, Ritter E, Paulo MJ, Cara N, Tacke E, Hofferbert H-R, Lübeck J, Strahwald J, Gebhardt C (2017) Physical mapping of QTL for tuber yield, starch content and starch yield in tetraploid potato (*Solanum tuberosum* L.) by means of genome wide genotyping by sequencing and the 8.3 K SolCAP SNP array. BMC Genomics 18:642. doi:https://10.1186/s12864-017-3979-9

Scott MF, Fradgley N, Bentley AR, Brabbs T, Corke F, Gardner KA, Horsnell R, Howell P, Ladejobi O, Mackay IJ, Mott R, Cockram J (2021) Limited haplotype diversity underlies polygenic trait architecture across 70 years of wheat breeding. Genome Biology 22 (1):137. doi:10.1186/s13059-021-02354-7

Sehgal D, Rosyara U, Mondal S, Singh R, Poland J, Dreisigacker S (2020) Incorporating Genome-Wide Association Mapping Results Into Genomic Prediction Models for Grain Yield and Yield Stability in CIMMYT Spring Bread Wheat. Frontiers in Plant Science 11. doi:10.3389/fpls.2020.00197

Selga C, Koc A, Chawade A, Ortiz R (2021a) A Bioinformatics Pipeline to Identify a Subset of SNPs for Genomics-Assisted Potato Breeding. Plants 10 (1):30

Selga C, Koc A, Chawade A, Ortiz R (2021b) A Bioinformatics Pipeline to Identify a Subset of SNPs for Genomics-Assisted Potato Breeding. Plants 10:30. doi:10.3390/plants10010030

Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W, Carboni MF, D'Ambrosio JM, de la Cruz G, Di Genova A, Douches DS, Eguiluz M, Guo X, Guzman F, Hackett CA, Hamilton JP, Li G, Li Y, Lozano R, Maass A, Marshall D, Martinez D, McLean K, Mejía N, Milne L, Munive S, Nagy I, Ponce O, Ramirez M, Simon R, Thomson SJ, Torres Y, Waugh R, Zhang Z, Huang S, Visser RGF, Bachem CWB, Sagredo B, Feingold SE, Orjeda G, Veilleux RE, Bonierbale M, Jacobs JME, Milbourne D, Martin DMA, Bryan GJ (2013) Construction of

Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps. G3 Genes|Genomes|Genetics 3:2031-2047. doi:10.1534/g3.113.007153

Sharma SK, Bryan GJ (2017) Genome sequence-based marker development and genotyping in potato. The potato genome:307-326

Sharma SK, MacKenzie K, McLean K, Dale F, Daniels S, Bryan GJ (2018) Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. G3 Genes|Genomes|Genetics 8:3185-3202. doi:10.1534/g3.118.200377

Shaw PD, Graham M, Kennedy J, Milne I, Marshall DF (2014) Helium: visualization of large scale plant pedigrees. BMC Bioinformatics 15 (1):259. doi:10.1186/1471-2105-15-259

Slater AT, Cogan NO, Forster JW, Hayes BJ, Daetwyler HD (2016) Improving genetic gain with genomic selection in autotetraploid potato. The plant genome 9:plantgenome2016.2002.0021. doi:10.3835/plantgenome2016.02.0021

Slater AT, Cogan NOI, Hayes BJ, Schultz L, Dale MFB, Bryan GJ, Forster JW (2014) Improving breeding efficiency in potato using molecular and quantitative genetics. Theoretical and Applied Genetics 127:2279-2292. doi:10.1007/s00122-014-2386-8

ŚLiwka J, Wasilewicz-Flis I, Jakuczun H, Gebhardt C (2008) Tagging quantitative trait loci for dormancy, tuber shape, regularity of tuber shape, eye depth and flesh colour in diploid potato originated from six Solanum species. Plant Breeding 127:49-55. doi:https://doi.org/10.1111/j.1439-0523.2008.01420.x

Song L, Endelman JB (2023) Using haplotype and QTL analysis to fix favorable alleles in diploid potato breeding. The Plant Genome 16:e20339. doi:https://doi.org/10.1002/tpg2.20339

Sood S, Bhardwaj V, Chourasia KN, Kaur RP, Kumar V, Kumar R, Sundaresha S, Bohar R, Garcia-Oliveira AL, Singh R (2022) KASP markers validation for late blight, PCN and PVY resistance in a large germplasm collection of tetraploid potato (Solanum tuberosum L.). Scientia Horticulturae 295:110859

Sood S, Bhardwaj V, Mangal V, Kardile H, Dipta B, Kumar A, Singh B, Siddappa S, Sharma AK, Dalamu, Buckseth T, Chaudhary B, Kumar V, Pandey NK (2024) Development of near homozygous lines for diploid hybrid TPS breeding in potatoes. Heliyon 10 (10). doi:10.1016/j.heliyon.2024.e31507

Sorensen PL, Christensen G, Karki HS, Endelman JB (2023) A KASP marker for the potato late blight resistance gene RB/Rpi-blb1. American Journal of Potato Research 100 (3):240-246

Stark JC, Thornton M, Nolte P (2020) Potato production systems. 1 edn. Springer Nature, Switzerland

Stich B, Gebhardt C (2011) Detection of epistatic interactions in association mapping populations: an example from tetraploid potato. Heredity 107:537-547. doi:10.1038/hdy.2011.40

Stich B, Van Inghelandt D (2018) Prospects and Potential Uses of Genomic Prediction of Key Performance Traits in Tetraploid Potato. Frontiers in Plant Science 9:159. doi:10.3389/fpls.2018.00159

Stokstad E (2019) The new potato. Science 363 (6427):574-577. doi:10.1126/science.363.6427.574

Sverrisdóttir E, Byrne S, Sundmark EHR, Johnsen HØ, Kirk HG, Asp T, Janss L, Nielsen KL (2017) Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. Theoretical and Applied Genetics 130:2091-2108. doi:10.1007/s00122-017-2944-y

Sverrisdóttir E, Sundmark EHR, Johnsen HØ, Kirk HG, Asp T, Janss L, Bryan G, Nielsen KL (2018) The Value of Expanding the Training Population to Improve Genomic Selection Models in Tetraploid Potato. Frontiers in Plant Science 9:1118. doi:10.3389/fpls.2018.01118

Tang D, Jia Y, Zhang J, Li H, Cheng L, Wang P, Bao Z, Liu Z, Feng S, Zhu X, Li D, Zhu G, Wang H, Zhou Y, Zhou Y, Bryan GJ, Buell CR, Zhang C, Huang S (2022) Genome evolution and diversity of wild and cultivated potatoes. Nature 606:535-541. doi:10.1038/s41586-022-04822-x

Tang X, de Boer JM, van Eck HJ, Bachem C, Visser RGF, de Jong H (2009) Assignment of genetic linkage maps to diploid Solanum tuberosum pachytene chromosomes by BAC-FISH technology. Chromosome Research 17:899-915. doi:10.1007/s10577-009-9077-3

Thérèse Navarro A, Tumino G, Voorrips RE, Arens P, Smulders MJM, van de Weg E, Maliepaard C (2022) Multiallelic models for QTL mapping in diverse polyploid populations. BMC Bioinformatics 23:67. doi:https://10.1186/s12859-022-04607-z

Thorup T, Tanyolac B, Livingstone K, Popovsky S, Paran I, Jahn M (2000) Candidate gene analysis of organ pigmentation loci in the Solanaceae. Proceedings of the National Academy of Sciences 97:11192-11197. doi:https://doi.org/10.1073/pnas.97.21.11192

Tinker NA, Bekele WA, Hattori J (2016) Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. G3 Genes|Genomes|Genetics 6:857-863. doi:10.1534/g3.115.024596

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

Uitdewilligen J, Wolters A, van Eck H, Visser R (2022) Allelic variation for alpha-Glucan Water Dikinase is associated with starch phosphate content in tetraploid potato. Plant Molecular Biology 108:1-12. doi:10.1007/s11103-021-01236-7

Uitdewilligen JG, Wolters A-MA, D'hoop BB, Borm TJ, Visser RG, Van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PloS one 8:e62355. doi:10.1371/journal.pone.0062355

Uitdewilligen JG, Wolters A-MA, D'hoop BB, Borm TJ, Visser RG, van Eck HJ (2015) Correction: A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PloS one 10:e0141940. doi:https://doi.org/10.1371/journal.pone.0062355

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Research 40:e115-e115. doi:10.1093/nar/gks596

Urbany C, Stich B, Schmidt L, Simon L, Berding H, Junghans H, Niehoff K-H, Braun A, Tacke E, Hofferbert H-R (2011) Association genetics in Solanum tuberosum provides new insights into potato tuber bruising and enzymatic tissue discoloration. BMC genomics 12:1-14. doi:https://doi.org/10.1186/1471-2164-12-7

van Eck HJ, Jacobs JME, Stam P, Ton J, Stiekema WJ, Jacobsen E (1994a) Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics 137:303-309. doi:https://10.1093/genetics/137.1.303

van Eck HJ, Jacobs JME, van den Berg PMMM, Stiekema WJ, Jacobsen E (1994b) The inheritance of anthocyanin pigmentation in potato (Solanum tuberosum L.) and mapping of tuber skin colour loci using RFLPs. Heredity 73 (4):410-421. doi:10.1038/hdy.1994.189

van Eck HJ, Oortwijn ME, Terpstra IR, van Lieshout NH, van der Knaap E, Willemsen JH, Bachem CW (2022) Engineering of tuber shape in potato (*Solanum tuberosum*) with marker assisted breeding or genetic modification using *StOFP20*. Research Square. doi:https://doi.org/10.21203/rs.3.rs-1807189/v1

van Eck HJ, Vos PG, Valkonen JPT, Uitdewilligen JGAML, Lensing H, de Vetten N, Visser RGF (2017) Graphical genotyping as a method to map Ny(o,n)stoand Gpa5 using a reference panel of tetraploid potato cultivars. Theoretical and Applied Genetics 130:515-528. doi:10.1007/s00122-016-2831-y

van Lieshout N, van der Burgt A, de Vries ME, ter Maat M, Eickholt D, Esselink D, van Kaauwen MPW, Kodde LP, Visser RGF, Lindhout P, Finkers R (2020) Solyntus, the New Highly Contiguous Reference Genome for Potato (*Solanum tuberosum*). G3 Genes|Genomes|Genetics 10:3489-3495. doi:https://10.1534/g3.120.401550

van Os H, Andrzejewski S, Bakker E, Barrena I, Bryan GJ, Caromel B, Ghareeb B, Isidore E, de Jong W, van Koert P, Lefebvre Vr, Milbourne D, Ritter E, van der Voort JNAMR, Rousselle-Bourgeois Fo, van Vliet J, Waugh R, Visser RGF, Bakker J, van Eck HJ (2006) Construction of a 10,000-Marker Ultradense Genetic Recombination Map of Potato: Providing a Framework for Accelerated Gene Isolation and a Genomewide Physical Map. Genetics 173 (2):1075-1087. doi:10.1534/genetics.106.055871

VanRaden PM (2008) Efficient methods to compute genomic predictions. Journal of dairy science 91:4414-4423. doi:10.3168/jds.2007-0980

Veeckman E, Van Glabeke S, Haegeman A, Muylle H, van Parijs FRD, Byrne SL, Asp T, Studer B, Rohde A, Roldán-Ruiz I, Vandepoele K, Ruttink T (2019) Overcoming challenges in variant calling: exploring sequence diversity in candidate genes for plant development in perennial ryegrass (Lolium perenne). DNA Research 26:1-12. doi:10.1093/dnares/dsy033

Vexler L, Leyva-Pérez MdlO, Konkolewska A, Clot CR, Byrne S, Griffin D, Ruttink T, Hutten RCB, Engelen C, Visser RGF, Prigge V, Wagener S, Lairy-Joly G, Driesprong J-D, Riis Sundmark EH, Rookmaker ANO, van Eck HJ, Milbourne D (2024) QTL discovery for agronomic and quality traits in diploid potato clones using PotatoMASH amplicon sequencing. G3 Genes|Genomes|Genetics 14:jkae164. doi:10.1093/g3journal/jkae164

Villano C, Miraglia V, Iorizzo M, Aversano R, Carputo D (2015) Combined Use of Molecular Markers and High-Resolution Melting (HRM) to Assess Chromosome Dosage in Potato Hybrids. Journal of Heredity 107. doi:10.1093/jhered/esv094

Vos P, Hogers R, Bleeker M, Reijans M, Lee Tvd, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research 23 (21):4407-4414. doi:10.1093/nar/23.21.4407

Vos PG, Paulo MJ, Bourke PM, Maliepaard CA, van Eeuwijk FA, Visser RGF, van Eck HJ (2022) GWAS in tetraploid potato: identification and validation of SNP markers associated with glycoalkaloid content. Molecular Breeding 42:76. doi:https://10.1007/s11032-022-01344-2

Vos PG, Paulo MJ, Voorrips RE, Visser RG, van Eck HJ, van Eeuwijk FA (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. Theoretical and Applied Genetics 130:123-135. doi:10.1007/s00122-016-2798-8

Vos PG, Uitdewilligen JG, Voorrips RE, Visser RG, van Eck HJ (2015) Development and analysis of a 20K SNP array for potato (Solanum tuberosum): an insight into the breeding history. Theoretical and Applied Genetics 128:2387-2401. doi:10.1007/s00122-015-2593-y

Wang F, Zou M, Zhao L, Xia Z, Wang J (2021) Genome-Wide Association Mapping of Late Blight Tolerance Trait in Potato (*Solanum tuberosum* L.). Frontiers in Genetics 12. doi:https://10.3389/fgene.2021.714575

Weber SE, Frisch M, Snowdon RJ, Voss-Fels KP (2023) Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. Frontiers in Plant Science 14:1217589

Werij JS, Kloosterman B, Celis-Gamboa C, De Vos CR, America T, Visser RG, Bachem CW (2007) Unravelling enzymatic discoloration in potato through a combined approach of candidate genes, QTL, and expression analysis. Theoretical and Applied Genetics 115:245-252. doi:https://doi.org/10.1007/s00122-007-0560-y

Werner J, Peloquin S (1990) Inheritance and two medchanisms of 2n egg fromation in 2x Potatoes. Journal of Heredity 81 (5):371-374

Wickland DP, Battu G, Hudson KA, Diers BW, Hudson ME (2017) A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC Bioinformatics 18:586. doi:10.1186/s12859-017-2000-6

Willemsen J (2018) The Identification of Allelic Variation in Potato. Ph.D. Thesis, Wageningen University and Research,Wageningen, The Netherlands,

Wilson S, Zheng C, Maliepaard C, Mulder HA, Visser RGF, van der Burgt A, van Eeuwijk F (2021) Understanding the Effectiveness of Genomic Prediction in Tetraploid Potato. Frontiers in Plant Science 12:672417. doi:10.3389/fpls.2021.672417

Wolters A-MA, Uitdewilligen JGAML, Kloosterman BA, Hutten RCB, Visser RGF, van Eck HJ (2010) Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. Plant Molecular Biology 73:659-671. doi:10.1007/s11103-010-9647-y

Won S, Park J-E, Son J-H, Lee S-H, Park BH, Park M, Park W-C, Chai H-H, Kim H, Lee J, Lim D (2020) Genomic Prediction Accuracy Using Haplotypes Defined by Size and Hierarchical Clustering Based on Linkage Disequilibrium. Frontiers in Genetics 11:134. doi:10.3389/fgene.2020.00134

Wu Y, Li D, Hu Y, Li H, Ramstein GP, Zhou S, Zhang X, Bao Z, Zhang Y, Song B, Zhou Y, Zhou Y, Gagnon E, Särkinen T, Knapp S, Zhang C, Städler T, Buckler ES, Huang S (2023) Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. Cell 186 (11):2313-2328.e2315. doi:10.1016/j.cell.2023.04.008

Yang H, Liao Q, Ma L, Luo W, Xiong X, Luo Y, Yang X, Du C, He Y, Li X, Gao D, Xue X, Shang Y (2021) Features and genetic basis of chlorogenic acid formation in diploid potatoes. Food Chemistry: Molecular Sciences 3:100039. doi:https://doi.org/10.1016/j.fochms.2021.100039

Yuan J, Bizimungu B, De Koeyer D, Rosyara U, Wen Z, Lagüe M (2020) Genome-Wide Association Study of Resistance to Potato Common Scab. Potato Research 63:253-266. doi:10.1007/s11540-019-09437-w

Zhang C, Wang P, Tang D, Yang Z, Lu F, Qi J, Tawari NR, Shang Y, Li C, Huang S (2019) The genetic basis of inbreeding depression in potato. Nature Genetics 51 (3):374-378. doi:10.1038/s41588-018-0319-1

Zhang C, Yang Z, Tang D, Zhu Y, Wang P, Li D, Zhu G, Xiong X, Shang Y, Li C, Huang S (2021) Genome design of hybrid potato. Cell 184:3873-3883.e3812. doi:https://doi.org/10.1016/j.cell.2021.06.006

Zhang F, Qu L, Gu Y, Xu Z-H, Xue H-W (2022a) Resequencing and genome-wide association studies of autotetraploid potato. Molecular Horticulture 2:6. doi:https://10.1186/s43897-022-00027-y

Zhang J, Yin J, Luo J, Tang D, Zhu X, Wang J, Liu Z, Wang P, Zhong Y, Liu C, Li C, Chen S, Huang S (2022b) Construction of homozygous diploid potato through maternal haploid induction. aBIOTECH 3:163-168. doi:https://10.1007/s42994-022-00080-7

Zhao D, Sandercock AM, Mejia-Guerra MK, Mollinari M, Heller-Uszynska K, Wadl PA, Webster SA, Beil CT, Sheehan MJ (2024a) A Public Mid-Density Genotyping Platform for Hexaploid Sweetpotato (Ipomoea batatas [L.] Lam). Genes 15 (8). doi:10.3390/genes15081047

Zhao D, Sapkota M, Glaubitz J, Bassil N, Mengist M, Iorizzo M, Heller-Uszynska K, Mollinari M, Beil CT, Sheehan M (2024b) A public mid-density genotyping platform for cultivated blueberry (Vaccinium spp.). Genetic Resources 5 (9):36-44. doi:10.46265/genresj.WQZS1824

Zheng C, Amadeu RR, Munoz PR, Endelman JB (2021) Haplotype reconstruction in connected tetraploid F1 populations. Genetics 219:iyab106. doi:10.1093/genetics/iyab106

Zhou Q, Tang D, Huang W, Yang Z, Zhang Y, Hamilton JP, Visser RGF, Bachem CWB, Robin Buell C, Zhang Z, Zhang C, Huang S (2020) Haplotype-resolved genome analyses of a heterozygous diploid potato. Nature Genetics 52:1018-1023. doi:10.1038/s41588-020-0699-x

Zhou Z, Plauborg F, Kristensen K, Andersen MN (2017) Dry matter production, radiation interception and radiation use efficiency of potato in response to temperature and nitrogen application regimes. Agricultural and Forest Meteorology 232:595-605. doi:10.1016/j.agrformet.2016.10.017

Zorrilla C, Navarro F, Vega-Semorile S, Palta J (2021) QTL for pitted scab, hollow heart, and tuber calcium identified in a tetraploid population of potato derived from an Atlantic× Superior cross. Crop Science 61 (3):1630-1651

Summary

In the current era of ever-advancing genomic technologies, the objective of “breeding by design”, the precise ability to combine favourable alleles across the genome, remains both an aspirational goal and a persistent challenge. In the context of potato, this vision has proven particularly elusive. The tetraploid nature of potatoes, coupled with high heterozygosity, and polygenic trait architecture complicates the prediction of genetic gain, while the cost of genome-wide genotyping has limited the routine use of genomic tools in breeding programs.

This thesis set out to ask a simple but ambitious question: is it possible to develop a genotyping system that is both cost-effective and broadly applicable, one that facilitates real-world breeding decisions, encompassing trait mapping, inbreeding analysis, and genomic prediction? To answer this question, I aimed to evaluate the strengths and limitations of such a system using realistic experimental material from the breeding programs of TEAGASC and the Fixation-Restitution project.

The outcome of this investigation was PotatoMASH: a low-cost, amplicon-based platform capable of generating both SNPs and read-backed haplotypes (haplotags). Designed with versatility as a priority, PotatoMASH supports multiple breeding applications from a single dataset.

The subsequent chapters examine how this platform can support key activities in genomics-assisted breeding, with particular emphasis on its application in Fixation-Restitution (Fix-Res) breeding, a strategy where both allele fixation and the preservation of strategic heterozygosity are essential to success.

Chapter 2 introduces PotatoMASH (Potato Multi-Allele Scanning Haplotags), a novel, low-cost, amplicon-based genotyping platform built on the “Genotyping in Thousands by Sequencing” framework. PotatoMASH targets 339 multi-allelic loci evenly spaced at 1 Mb intervals throughout the euchromatic portion of the genome and uses deep sequencing. In conjunction with Stack Mapping Anchor Points (SMAP) software for read-backed haplotyping, PotatoMASH simultaneously generates two types of markers: SNPs and short-read multi-allelic haplotypes (haplotags). The platform was validated across more than 700 potato clones. It reliably captured high levels of haplotypic diversity, ranging from 2 to 14 haplotags per locus, and proved useful across a variety of breeding applications. These included the detection of diagnostic markers for pest resistance, genome-wide association studies (GWAS) and genetic mapping in a biparental population. At a cost of €4-5 per sample (excluding bioinformatics and labour costs), PotatoMASH offers a flexible and scalable genotyping solution that combines affordability with rich allelic information. Its development marks an important step toward making genome-wide marker data accessible for routine use in breeding programs, laying the technical foundation for the downstream analyses presented in later chapters.

Chapter 3 explores the use of PotatoMASH to discover quantitative trait loci (QTL) in a diverse panel of 618 diploid potato clones that were part of the Fixation-Restitution breeding project. This panel was phenotyped for 23 agronomic, morphological, and quality traits over three years as part of a collaborative effort involving six European diploid breeding programs. GWAS were performed for all traits using both SNPs and haplotags, derived from read-backed phasing of SNP combinations. High haplotypic diversity was observed in the panel, with 2 to 30 haplotags per locus. A total of 37 QTL were identified across 20 traits: 10 detected by both marker types, 14 uniquely by haplotags, and 13 by SNPs. Although haplotags were initially hypothesized to outperform SNPs due to their multi-allelic nature, the two marker types captured different signals. Further investigation revealed that haplotag-only QTL often involved SNPs embedded in non-significant haplotags, whereas SNP-only QTL typically involved markers distributed across several low-frequency haplotags, reducing the power of haplotype-based tests. These findings highlight the complementary value of using both marker types in parallel.

This chapter presents one of the most extensive multi-trait association studies in diploid potato to date. We report 19 novel QTL across nine traits: skin smoothness, sprout dormancy, total tuber number, tuber length, yield, chipping colour, after-cooking blackening, cooking type, and eye depth, mapped across eight chromosomes.

Chapter 4 investigates the use of haplotag-based genotyping to track homozygosity changes, estimate selfing rates, and monitor genome-wide residual heterozygosity (RH) in diploid potato breeding populations. A collection of 271 inbred diploid clones from the Fixation-Restitution breeding program of Plant Breeding Wageningen University & Research was genotyped using PotatoMASH to provide a snapshot of the genetic composition shaped by over 40 years of breeding. In addition, 174 individuals from a self-compatible lineage, spanning three generations of inbreeding, were analysed to identify chromosome-wide patterns of allele fixation and residual heterozygosity. Homozygosity estimates derived from haplotags were consistently more accurate than those based on SNPs. For example, in the S1 generation, SNPs estimated 83.1% homozygosity, while haplotags gave a more accurate and conservative estimate of 59.1%, indicating that SNP-based dosage scoring significantly overestimate homozygosity, particularly in highly heterozygous material or in the absence of parental genotypes. Despite progressive selfing, several genomic regions remained consistently heterozygous. Chromosome 5 showed complete retention of heterozygosity across all generations. Additional resistant loci included C3_8 (heterozygous in 100% of individuals) on chromosome 3 and C12_10 (heterozygous in over 90%) on chromosome 12. These regions co-located with previously reported QTL associated with fertility and performance traits known to be affected by inbreeding depression, suggesting selective retention of heterozygosity at functionally important loci. This chapter demonstrates that haplotype-based genotyping provides a robust and high-resolution approach to tracking inbreeding and managing heterozygosity in diploid breeding programs. In the context of Fix-Res breeding, where strategic heterozygosity at key loci may be critical to success, haplotags offer a powerful tool to guide selection and crossing decisions.

Chapter 5 evaluated the use of PotatoMASH as a low-cost genotyping platform for genomic prediction (GP). In a tetraploid population phenotyped for fry colour, PotatoMASH performance was compared to high-density GBS data (43.6k SNPs). Despite using far fewer markers (2,236 SNPs and 2,000–3,390 haplotags), predictive accuracy (PA) was reduced by only 14% with SNPs and 9% with haplotags relative to GBS, at a cost 10-20 times lower per clone. In a diploid breeding panel of 558 clones, GP models were developed for 23 agronomic, quality, and morphological traits, with predictive accuracies ranging from 0.29 to 0.81. Based on findings from previous chapters, where SNPs and haplotags showed distinct differences in performance for QTL detection, we expected similar contrasts in GP. However, the differences in predictive accuracy between the two marker types were generally subtle, even though some were statistically significant. Haplotags outperformed SNPs in 11 traits, while SNPs outperformed haplotags in six traits; the remaining six traits showed comparable results. In previous chapters, haplotags were constructed using the SMAP *haplotype-sites* module, which combines pre-called SNPs into haplotypes based on read-backed phasing. In this chapter, we also evaluated haplotypes generated using the newer SMAP *haplotype-window* module, which reconstructs haplotypes directly from full read sequences. Because *haplotype-window* does not rely on predefined SNPs, it captures broader genetic diversity. Additionally, it generates sequence-based identifiers that are consistent across populations, improving marker comparability and enabling cross-population interoperability for research and breeding applications.

Chapter 6 provides a synthesis of the work presented in this thesis, reflecting on the development and application of PotatoMASH within the broader context of breeding by design in potato. It revisits the technical and practical challenges that have historically limited the implementation of genomics-assisted breeding in this crop, namely, its polyploid nature, high heterozygosity, and the cost of genome-wide marker systems, and outlines how PotatoMASH was designed to address these limitations. The chapter draws on findings from earlier chapters to show how a single, low-cost genotyping platform can support key genomic tools, GWAS, MAS, genomic prediction, and inbreeding tracking, within a single-genotyping event, enabling breeders to reuse the same dataset across multiple decision points. These tools are discussed in relation to their roles within the Fix-Res framework, from pre-breeding and introgression through diploid selection and interploidy crosses. The advantages of read-backed haplotypes are emphasized throughout, particularly their ability to capture multi-allelic variation, provide improved resolution in marker–trait associations, produce more accurate homozygosity estimates, and track genetic segments through inbreeding. While the discussion is structured around the Fix-Res approach, the genotyping platform and analytical strategies described are also considered in the context of other breeding systems that rely on structured selection and genome-informed decision-making.

Conclusion: towards breeding-by-design with read-backed haplotypes

This thesis demonstrates that read-backed haplotypes**,** when derived from a carefully designed and cost-effective amplicon sequencing marker platform like PotatoMASH, can support breeding-by-design in potato. Each chapter contributes to this central aim:

- Chapter 2 established a novel genotyping platform.
- Chapter 3 validated its value for QTL discovery.
- Chapter 4 showed its application for tracking inbreeding.
- Chapter 5 proved its utility for genomic prediction.
- Chapter 6 unified these findings under a modern breeding framework.

Together, these chapters show that the multi-allelic nature of haplotags is a powerful complement to SNPs**,** offering a higher resolution and improved biological insight while maintaining affordability. PotatoMASH enables a unified genotyping strategy across breeding phases and is especially well-suited to Fix-Res breeding, but applicable more broadly to diploid and tetraploid systems alike. By integrating cost, resolution, and versatility, this work represents a meaningful step toward realizing practical genomic selection and allele fixation in potato breeding.

Acknowledgments

At the end of my PhD journey, I would like to thank all the people who helped me reach this point. Without your support along the way, I couldn't have done this alone.

First and foremost, I want to thank my supervisors, **Dan Milbourne** and **Herman van Eck**. **Dan**, you were my rock throughout this entire period. Thank you for your encouragement and unwavering support. You pushed me to learn and explore, showed genuine interest in my ideas, and your open-door policy made it easy to brainstorm and consult. These conversations helped me digest and understand complex new concepts. Alongside encouraging me to strive for excellence, you truly cared about my wellbeing. I felt that you always had my back, and you brought a fun and relaxed attitude that made working with you genuinely enjoyable. Your broad perspective, collaborative mindset, and strategic thinking taught me so much. **Herman**, thank you for sharing your deep knowledge of potato genetics with me, I learned a great deal from you. Your thorough and thoughtful approach to research taught me to question data and results critically, and your sharp eye for detail kept me on my toes, striving for the highest standards in my work. Although we only met in person a couple of times, I still remember our bike ride through Wageningen and your warm hospitality. Those moments stayed with me and meant a lot.

To my promotor, **Richard Visser**, your support throughout this journey made a real difference. Your constructive, to-the-point feedback and encouraging attitude were tremendously helpful in getting me across the finish line. At key moments, your clear guidance and quick responses helped me move forward with focus and confidence. I'm very grateful for your steady presence and for the practical and personal support that played such an important role in bringing this thesis to completion.

To two dear people who stood by me daily throughout this journey, **Maria de la O Leyva-Perez** and **Stephen Byrne**. **Mariola**, you were a postdoc when I started, and your role was to teach me the molecular and bioinformatics methods for PotatoMASH, which you developed. But over time, you became a dear friend. You were there to offer advice, discuss even the smallest details, support me through every failure, and celebrate each success. Your positivity, laughter, and belief in me gave me strength. Your scientific skills and creative thinking showed me that there is always a solution if we think outside the box. I admire you as a researcher and as a person. **Stephen**, I don't even know where to begin thanking you. You were a mentor, a friend, and a colleague, always willing to help, even without any formal obligation to me or the project. You offered ideas for tackling complex data, statistics, and bioinformatics, and were especially supportive during the writing process, patiently helping even with the earliest drafts. I learned so much from you about scientific writing and how to approach it. You are the kind of scientist I aspire to be, professional, thoughtful, and generous.

To all the **potato team** in **Teagasc** thank you for introducing me to potatoes! Specifically to **Denis Griffin**, thank you for sharing your extensive potato knowledge, always taking the time to answer even the most basic questions. You are supportive and always have a kind word. Thank you as well for introducing me to people in the field, you are absolutely the best person to attend events and conferences with!

I would like to express my deep gratitude to the researchers and colleagues who collaborated with me and contributed significantly to the development of my skills. **Corentin Clot**, you were the first PhD student I interacted with, and the only one during my first six months, when I started remotely at the beginning of COVID. You introduced me to the Linux environment, walked me through my first-ever analysis of NGS data, helped me with university documents, and patiently explained the project and data. I'm very grateful for your help in those early days, and for your continued generosity in the years that followed. **Agnieszka Konkolewska**, thank you so much for your support and guidance in navigating R and statistics. I learned a lot from you about GWAS, Genomic Prediction, and data visualization. **Tom Ruttink**, thank you for all your help with SMAP and haplotyping, and for the detailed comments that really helped me move forward in analysis and writing.

Looking back, I feel fortunate to have done my PhD at **Crop Science Department**, **Teagasc Oak Park**, a uniquely inclusive and supportive environment. This became especially clear during a time when I faced health challenges and needed extra flexibility. As a foreign student, that period was particularly complex, and I'm truly grateful for the kindness and care I received. **Ewen Mullins** thank you for your support during my time as a students representative, where I saw firsthand how genuinely you care about students' wellbeing. I also personally felt your behind-the-scenes support over the past two years, which was incredibly reassuring and something I deeply appreciate. **Eleanor Butler**, thank you for always finding the time to help me with administrative issues, and for your professional sensitivity. **Helena Meally**, **Fiona Hutton** and **Colum Kennedy** thank you for helping me with molecular lab related and other lab matters. **Paul Cormican**, thank you for your patience and help while I was finding my footing in HPC and Linux.

To everyone in **Plant Breeding, Wageningen University and Research**, thank you for always being welcoming and helpful whenever we interacted. To the **EPS** coordinators, thank you for the clear communication and creating opportunities to include me as a sandwich PhD, and especially to **Susan Urbanus**, thank you so much for your caring approach, thoughtful and helpful advice at critical moments.

I also want to thank all the partners involved in the DIFFUGAT and Fixation-Restitution projects, especially the breeders who provided me with material and data and patiently endured all my potato-related questions! **Ronald Hutten** and **Christel Engelen** from **Wageningen**, **Vanessa Prigge** and **Silke Wagener** from **SaKa** and **Jan-David Driesprong** from **Meijer Potato**, **Gisele Lairy-Joly** from **Germicopa**, **Ea Høegh Riis Sundmark** from **Danespo**, **Nico Rookmaker** from **AVERIS Seeds**, thank you. **Vanessa** and **Ronald,** thank you for your guidance and availability to discuss potato breeding with me.

To my external advisor, **Amit Gur** at the Agricultural Research Organization (ARO) in Israel, you were my supervisor during my Master's and I am grateful for your ongoing support and belief in me throughout my PhD. Your advice was always grounding and insightful and your perspective helped me to make thoughtful decisions along the way, thank you!

To my dear friends, **Diana Bucur**, my office mate from day one, and quickly also my confidant and partner in everything, from trips and sports to everyday life. Thank you for your constant presence, your listening ear, and for being such a steady and supportive friend throughout this journey. **Rachel Keirse**, you were one of the first people I connected with when I arrived in Carlow during the pandemic, and your friendship has meant a lot to me. Thank you for introducing me to Irish culture, for showing me places I would never have discovered on my own, and for always making me smile with your sense of humor. My wonderful **Lucy Cases**, thank you for all the serious, and mostly hilarious, moments. **Francesca Messiti**, thank you for always being ready with a pint to celebrate and cherish the moment. **Manfred Klass**, thank you for introducing me to traditional Irish music and for many interesting conversations. Dear **Keith Ward**, thank you for keeping me on top of my lifting game, you helped me stay sane, strong, and smiling through the ups and downs of the past four years. **Rabisa Zia** and **Jie Huang**, thank you for your supportive friendship and the good times together. **Elena Grosu**, it was great being a student rep alongside you, I admire your dedication. To all my fellow Walsh Scholars and colleagues at Teagasc over the past four years, thank you for being such a supportive and fun group and for working together. Thanks to you all, these four years were filled with laughter and friendship, and I will carry you with me as lifelong friends.

,תודה למשפחתי האהובה, שעמדה לצידי לאורך כל הדרך - **לסבתי היקרה, לאימי, לאחיותיי ואחיי ובמיוחד לאחייניות ולאחיינים המתוקים: איילה, מילכה, מיכלי, יעלי, שוהם ואדל**. המכתבים והיצירות שהכנתם לי לקישוט המשרד שימחו אותי מאוד, והזכירו לי בכל יום את הדברים החשובים והמרגשים שיש בחיים.

English translation of Hebrew paragraph: Thank you to my beloved family, who stood by me every step of the way - to my dear grandmother, my mother, my sisters and brothers, and especially to my sweet nieces and nephews: Ayala, Milka, Michali, Yaeli, Shoham, and Adel. Your letters and drawings for my office brought me so much joy. They reminded me every day of the meaningful and heartwarming things in life.

Finally, I dedicate this thesis to Sharon Brennan.

Sharon, thank you, for the remarkable woman you are. You were always ready with a kind word and genuine interest. I especially appreciated how you helped me when I was sick: you cared so much, helped me navigate Irish healthcare bureaucracy, picked up the phone to get answers for me, and went out of your way to support me. You were truly happy for me when I finally got the help I needed. I will always carry your loving memory in my heart.

About the Author



Lea Vexler was born and raised in Israel. She earned her B.Sc. in Chemistry and Biology from the Hebrew University of Jerusalem. During that time, she developed a strong interest in plant science, and particularly in plant breeding. She continued her academic journey with an M.Sc. in Field and Vegetable Crops at the Volcani Institute (Neve Ya'ar, Israel), in collaboration with the Faculty of Agriculture of the Hebrew University in Rehovot. Her master's research, under the supervision of Dr. Amit Gur and Dr. Yaakov Tadmor, focused over the course of three years on the characterization of high β-carotene content in watermelon.

Following her M.Sc., Lea worked at Equinom, a plant breeding company in Israel, where she focused on breeding legumes for high protein content. Over the course of her academic and professional experience, she developed a deep appreciation for the importance of making genomic breeding technologies both accessible and cost-effective, particularly for small and medium-sized breeding companies, to facilitate faster and more efficient crop improvement.

Motivated by this goal, Lea began her PhD research at Teagasc, Ireland, under the supervision of Dr. Dan Milbourne, in collaboration with the Plant Breeding group at Wageningen University & Research, the Netherlands, under the supervision of Dr. Herman van Eck and Prof. Richard Visser. Her PhD research in potato breeding and genetics focused on developing genomic tools for use in practical breeding. The findings of her doctoral work are documented in this thesis. Lea is currently working as a postdoctoral research officer in the IPMorama project (https://ipmorama.eu/) at Teagasc, where she continues to develop and apply genomic technologies in potato breeding.

Authorship Statement

Chapter 1 – *General Introduction*. I proposed the layout and structure of this chapter and wrote the first draft. Based on discussions with colleagues and my co-promotors, I revised the early drafts. I received feedback from my promotor and co-promotors and revised the chapter accordingly to produce the final version.

Chapter 2 – *PotatoMASH: A low-cost marker platform for genomic analysis in potato.* I contributed to optimization and the protocol publication of molecular and bioinformatics pipelines, preparing data for downstream application, data analysis, visualisation and interpretation of the results in collaboration with the first co-author who developed of the platform. I contributed to drafting the manuscript and finalising the published version.

Chapter 3 – *GWAS Using PotatoMASH Reveals QTL for Multiple Agronomic Traits in Diploid Potato.* I coordinated the data collection from all partners and participated in proposing the methodology. I isolated DNA, constructed PotatoMASH libraries, and performed data curation and statistical analysis for both phenotypic and genotypic datasets. I conducted the genome-wide association studies (GWAS), largely interpreted the results, and produced all figures and tables. Co-authors provided the germplasm and collected the phenotypical data. I drafted the first version of the manuscript and revised the text based on discussions and feedback from co-authors, my promotor, and co-promotors.

Chapter 4 – *Tracking Inbreeding and Heterozygosity Using Haplotype-Based Genotyping in Diploid Potato.* I contributed to the design of the research, proposed the methodology and research questions. I genotyped the panels, performed the bioinformatics and inbreeding analysis, largely interpreted the results, and produced the figures and tables. Co-authors provided the germplasm. I wrote the first draft of the manuscript and revised it based on feedback and discussions with co-authors, my promotor and co-promotors.

Chapter 5 – *Genomic Prediction Using PotatoMASH: A Cost-Effective Alternative to High-Density Genotyping in Potato.* I contributed to the design of the research, proposed the methodology and research questions. I performed bioinformatics and genomic prediction analyses, largely interpreted the results, and produced the figures and tables. I wrote the first draft of the manuscript and revised it based on feedback and discussions with co-authors, my promotor, and co-promotors.

Chapter 6 – *General Discussion.* I proposed the structure and content for this chapter, synthesizing key discussion points and future research directions. I drafted the initial version and revised it based on discussions and suggestions from my colleagues and co-promotors. After further feedback from my co-promotors and promotor I finalised this chapter.

In this thesis, I used ChatGPT (GPT-4o, OpenAI's large-scale language model, 2024) to improve the phrasing and clarity of my original drafts. All scientific content, analysis, and interpretation are my own.

The research described in this thesis was financially supported by the DIFFUGAT project ("Diploid Inbreds For Fixation, and Unreduced Gametes for Tetraploidy") via funding from the Department of Agriculture Food and the Marine (DAFM) under the ERA-NET Cofund SusCrop (Grant No. 771134), being part of FACCE-JPI, and the Dutch Topsector Horticulture & Starting Materials project "A new method for potato breeding: the 'Fixation-Restitution' approach"; (grant number TU18075).

Contributions were made by the potato breeding companies Meijer Potato; Danespo A/S; SaKa Pflanzenzucht GmbH & Co. KG; Germicopa Breeding and Averis Seeds B.V.

We acknowledge the support of the Teagasc/IPM Potato Group Breeding Program.

Lea Vexler was supported by a Teagasc Walsh Scholarship.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Lea Vexler using ChatGPT (GPT-4o, OpenAI's image generation model, 2024)

Printed by ProefschriftMaken


CCCATAGGTCCTCGCGGCGGGGGGGGC


ATATAGGGTAGGTCATATCTACTACG