agronomy

MDPI

Article

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

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] and Dan Milbourne [1,*]

1 Teagasc, Crop Science Department, Oak Park, R93 XE12 Carlow, Ireland
2 Plant Breeding Labortory, Wageningen University, 6700 AJ Wageningen, The Netherlands
3 Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), 9090 Melle, Belgium
4 Department of Mathematics and Statistics, University of Otago, Dunedin 9016, New Zealand
* Correspondence: maria.leyva@teagasc.ie (M.d.l.O.L.-P.); Dan.Milbourne@teagasc.ie (D.M.)

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.

Keywords: potato; genetics; breeding

Citation: Leyva-Pérez, M.d.l.O.; Vexler, L.; Byrne, S.; Clot, C.R.; Meade, F.; Griffin, D.; Ruttink, T.; Kang, J.; Milbourne, D. PotatoMASH—A Low Cost, Genome-Scanning Marker System for Use in Potato Genomics and Genetics Applications. *Agronomy* **2022**, *12*, 2461. https://doi.org/10.3390/agronomy12102461

Academic Editor: Jadwiga Śliwka

Received: 1 September 2022
Accepted: 5 October 2022
Published: 11 October 2022

1. 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. "Pre-designed" 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 [1,2]. 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 [3,4]. 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 [5]. Sequence capture-based GBS approaches have also proven to be powerful in potato, both to survey variation on a genome-wide level [6] or to target specific motifs such as resistance loci [7]. 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. [8] 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 [9,10]. 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 [11], 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. [6]. 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. 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 [6,12–14]. 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. referred to as "Tag-level haplotypes" or haplotags [15]. 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. 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 [15–17].

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 [11]. Because of the availability of a reference pseudochomosome molecule-scale assembly in potato, we utilised a novel algorithm called SMAP (Stack Mapping Anchor Points) [18], 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 [4] 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.

2. 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 [4],

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.

2.1. 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 [11,19,20].

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

Chr/arm	Start	End	Length	Core	Diagnostic	Total
			(Mb)	Loci	Loci	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

We mapped WGS re-sequencing data of 75 commercial cultivars ([21]; 33× coverage, 5 pools of 15 cultivars) to the *Solanum tuberosum* genome DM_v4.04 [22] using BWA-MEM [23]. We used Popoolation software [24] 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 [25] 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 [26] to check for sequence similarity with off-target regions (only single-copy regions were retained). Primer 3 plus [27] 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) [28] or retrieved from the literature [29]. A summary of the pipeline employed for primer design is illustrated in Figure 1.

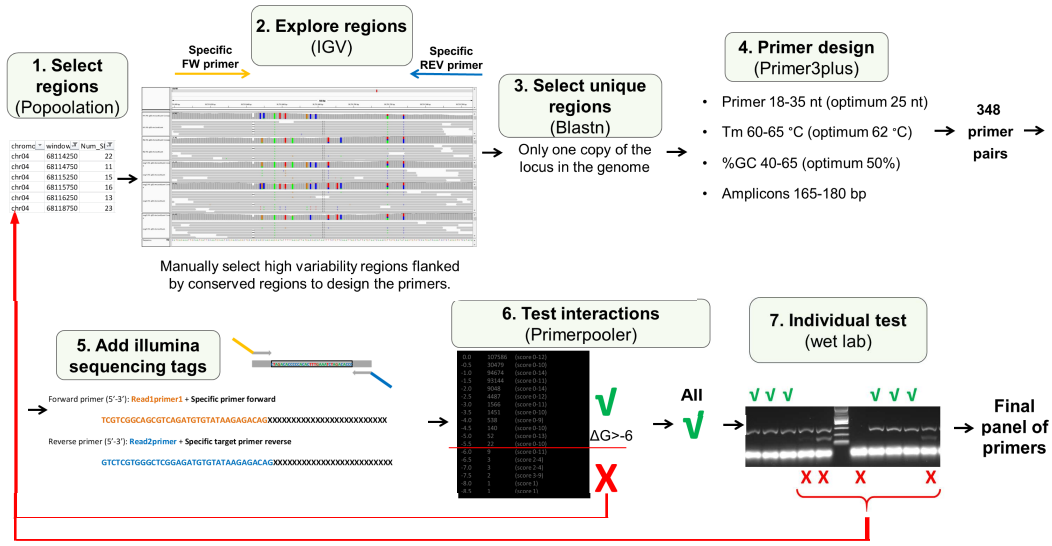


Figure 1. Pipeline for PotatoMASH primers design.

The method to construct multiplex amplicon libraries for Illumina sequencing was based on the GT-Seq method [8]. 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 5′ extreme of each forward primer (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-FW primer) and the tag sequence for R2 Illumina primer at 5′ end of each reverse primer (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-REV primer). Thus, the primers for the first PCR ranged from 51-69 nt in length.

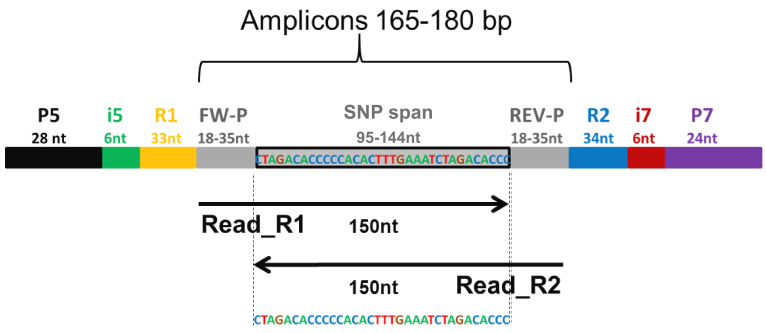


Figure 2. PotatoMASH library structure in relation to paired-end 150nt Illumina reads Read_R1 and Read_R2.

To avoid primer interactions and putative secondary products, all designed primers were tested in silico using primerpooler [30] 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 $MgCl_2$, 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 ddH_2O 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).

2.2. 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 [4]. 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.

2.3. 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 ddH_2O, 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 °C min); 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 ddH_2O 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.

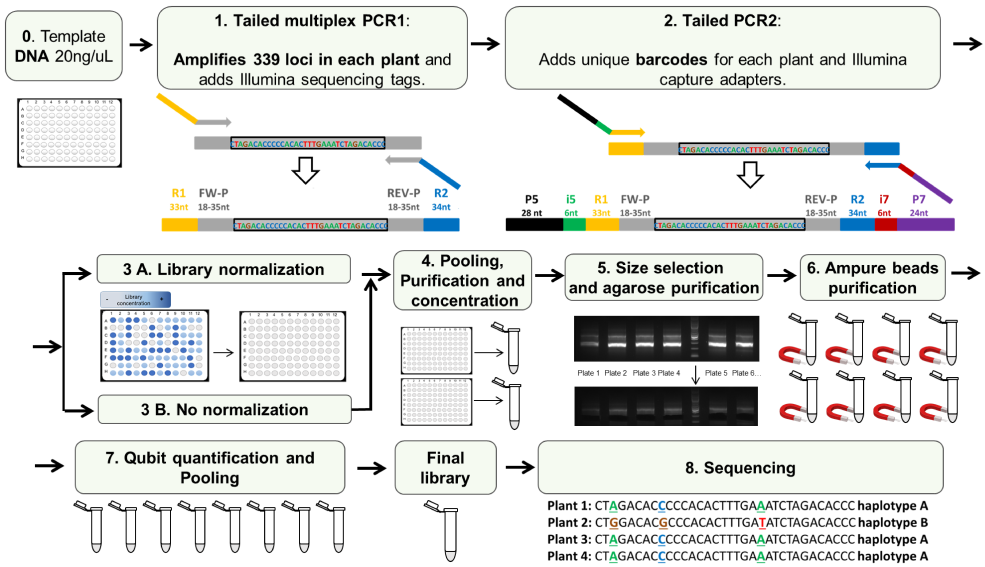


Figure 3. PotatoMASH library construction.

In contrast with the original GT-Seq protocol [8], 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 ddH_2O 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).

2.4. Multiallelic Haplotype Analysis

Fastq files were de-multiplexed, barcodes removed, and read pairs merged using FLASH [31] (min overlap -m 50; Max overlap -M 150). We filtered the merged reads using the fastx toolkit [32] (minimum base quality (-q30) in 90% of bases (-p90)). Merged and filtered reads were then mapped to the *S. tuberosum* genome v4.04 [22] with BWA-MEM [23]. Variant calling was performed with bcftools [33], 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 –max-missing 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 [18] 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.

2.5. 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 [4]. 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 [4]. GWAS was carried out with the R package GWASpoly [34]. 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).

2.6. 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 [35] and the breeding clone bearing a *S. microdontum* and *S. tuberosum* ancestry IVP10-281-1 [36]. 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 multi-allelic 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 [37]. 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 [38]. 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 [39] 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.

3. Results

3.1. 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.

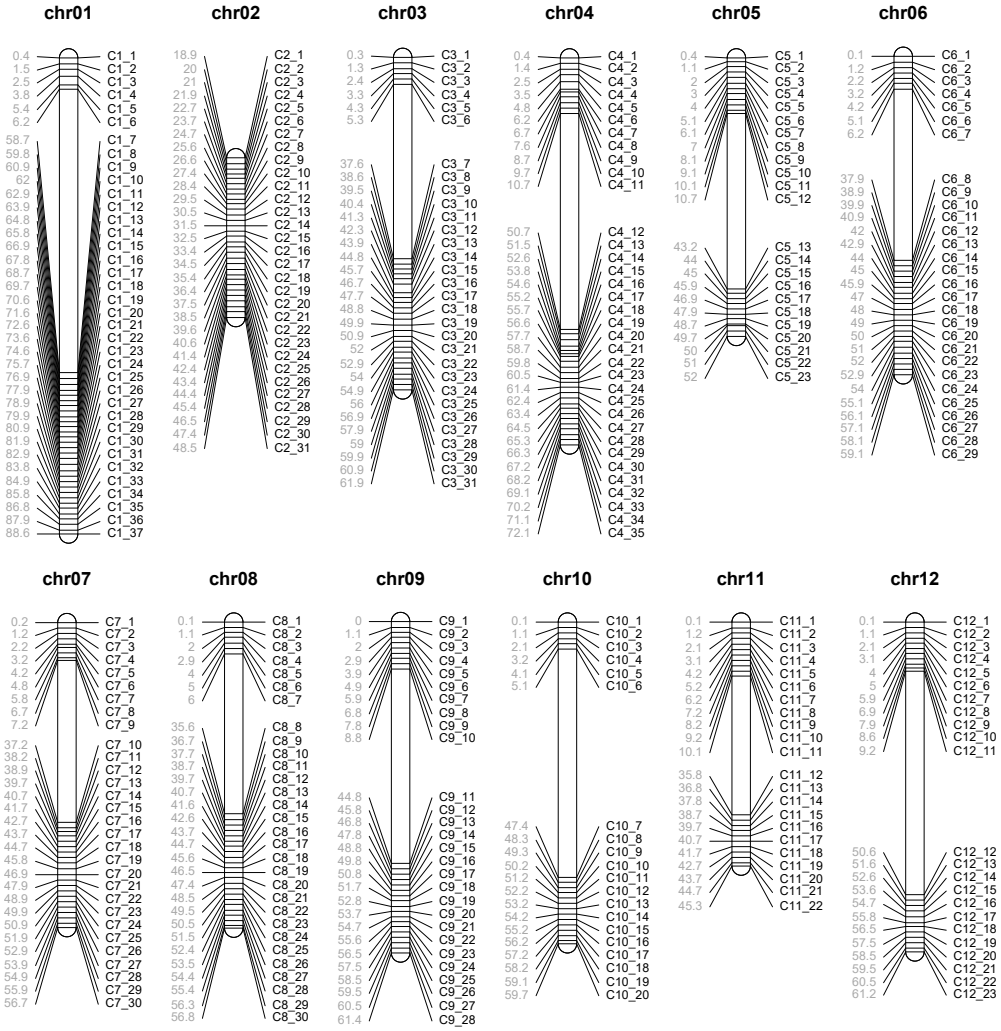


Figure 4. The physical map of potato based on pseudochromosome molecule assembly of the DM_v4.04 reference sequence. The positions of the 339 PotatoMASH core loci are represented in Mb intervals.

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.

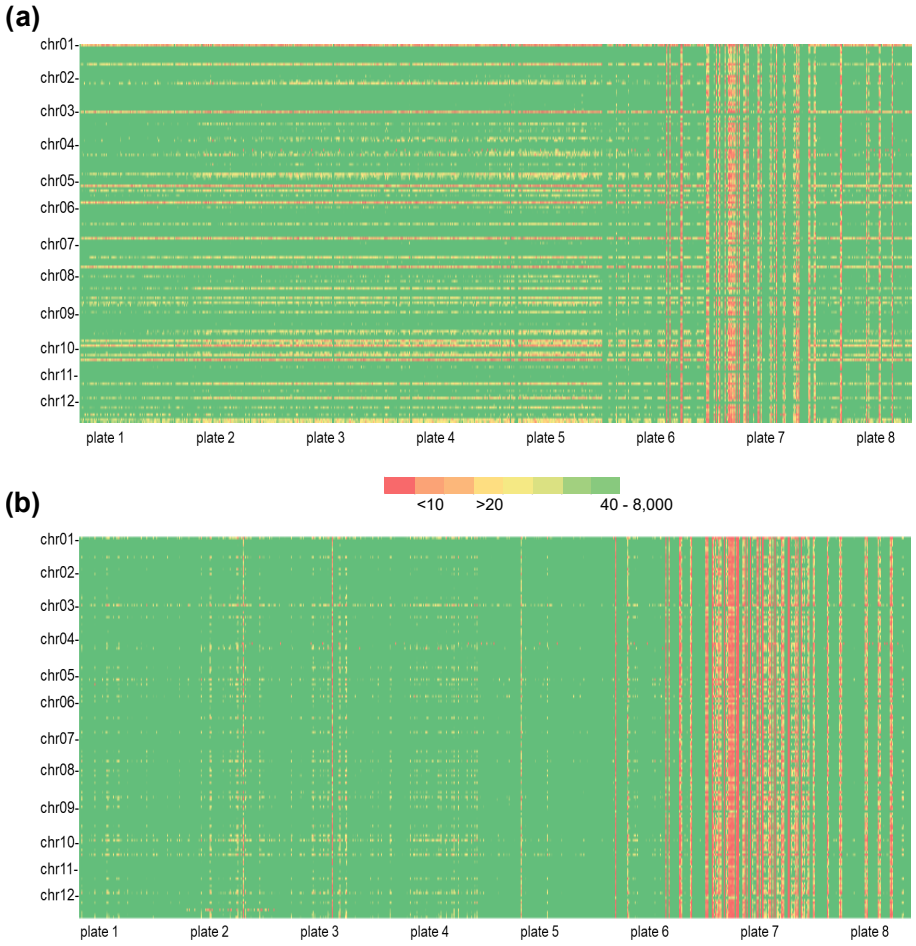
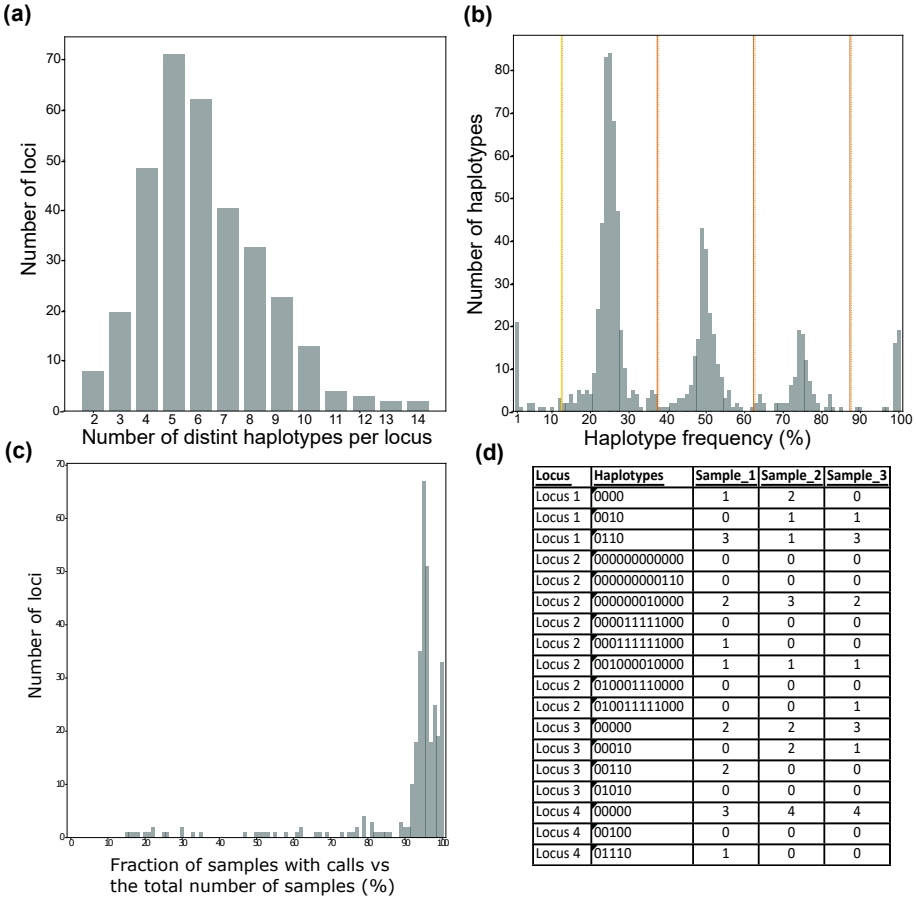


Figure 5. Coverage of the 339 PotatoMASH core loci. Heat map of the number of merged and filtered reads of 765 samples (in columns) that mapped to each locus (in rows). (**a**) Normalized library (**b**) Non-normalized library.

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 completeness; Distribution of samples across the locus set. (**d**) Example of tabular data generated by SMAP haplotype-sites with 3 samples, 4 loci, 18 haplotypes, and tetraploid discrete dosage calls for each locus/sample.

3.2. 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.
[29]	*S. scabies*	c2_17867	chr02:36548178 [T/C]	C2_B2	011110	99.9%
[29]	*S. scabies*	c2_17864	chr02:36550070 [T/C]	C2_B3	0011110	99.9%
TPBP	*G. pallida (Pa2/3)*	Gpa4	chr04:4782401 [A/G] chr04:6191864 [A/T]	C4_5	001110	99.9%
[21]	*P. infestans*	R2	chr04:6191873,76,77 [TGATT/CGAAA]	C4_6	Not detected	NA
[40]	*G. pallida*	Gpa5	chr05:5485534 [T/A]	C5_B9	0110101010111100101	96.5%
[21]	*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
[41]	PVY	Ny(o,n)sto	chr11:284162 [T/C] 68 [T/C]	C11_B1	000101010010	98.95%
[42]	*S. endobioticum*	Sen1	chr11:3928601 [A/G]	C11_B3	001100	100%
[43]	PVY	Ry-sfto	chr12:59957417 [G/A]	C12_B6	Not detected	NA

3.3. 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).

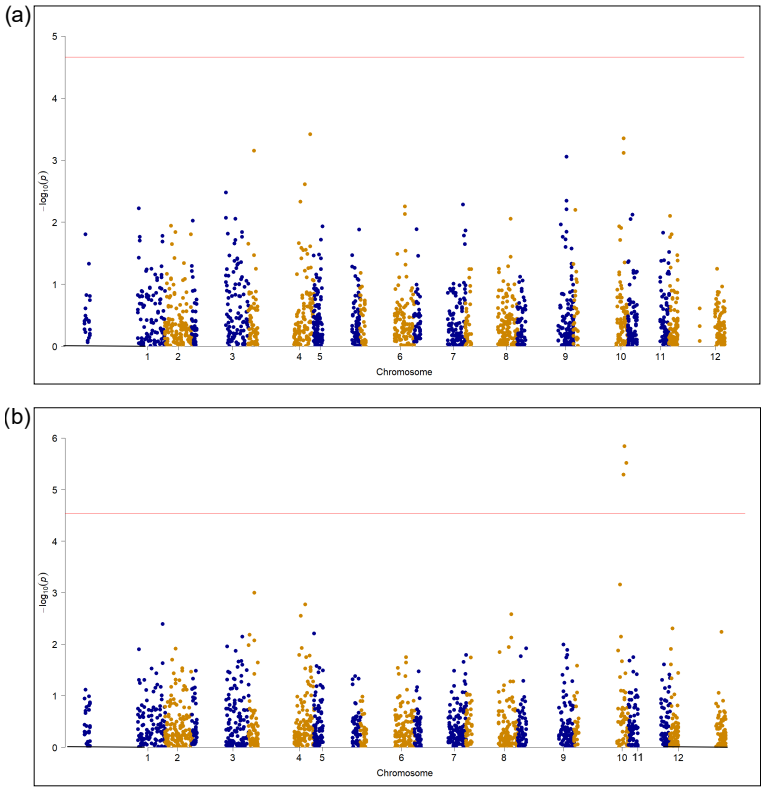


Figure 7. Manhattan plot of GWAS results, additive model, for 'off-the-field' fry colour (OTF) in 2017 population with the SNPs (**a**) and the multiallelic haplotypes (**b**). Horizontal line shows the QTL significance threshold at 4.54 (Bonferroni correction, level = 0.05).

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, C10_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. [4] 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.

3.4. Linkage Map Construction Using PotatoMASH Haplotypes

We also applied PotatoMASH to a bi-parental diploid mapping population (FRW19-112), 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 multi-allelic 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 multi-allelic 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.

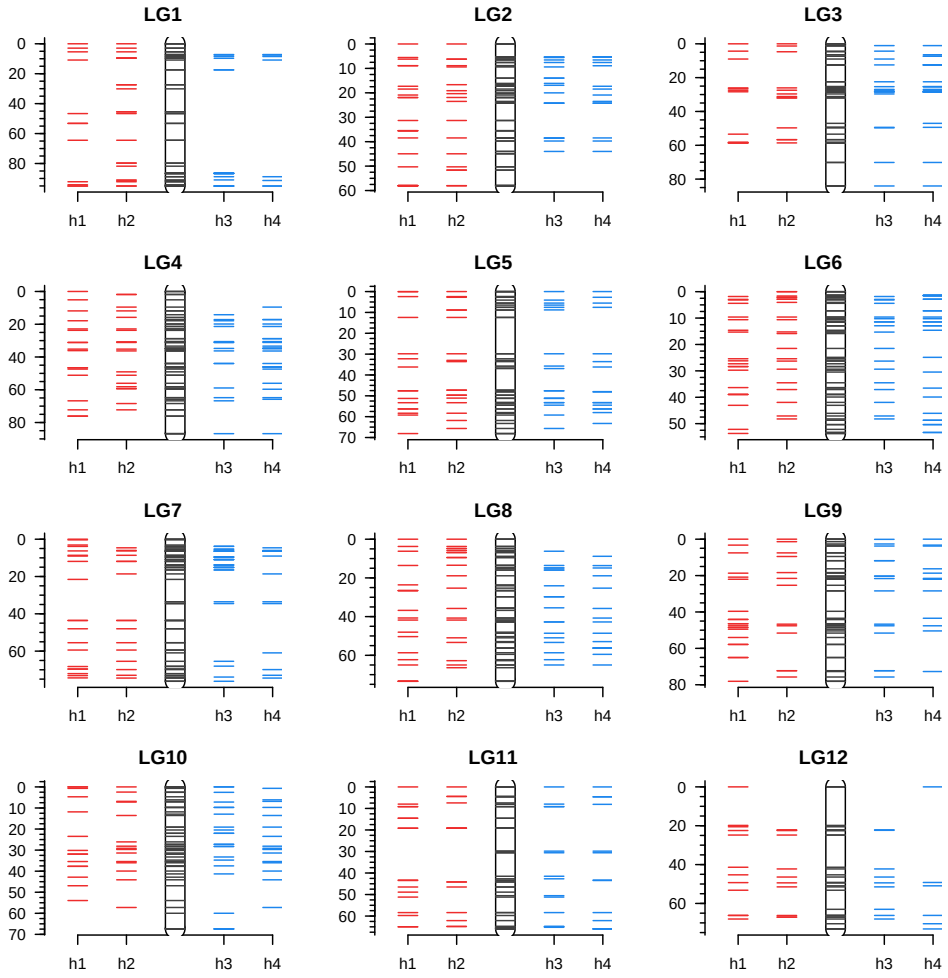


Figure 8. Phased homologue-specific map of population FRW19-112. Homologue maps from the diploid parent RH89-039-16 are shown in red (h1–h2), blue for IVP10-281-1 the diploid parent (h3–h4), and the integrated chromosomal map is shown in black.

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 [44]. 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.

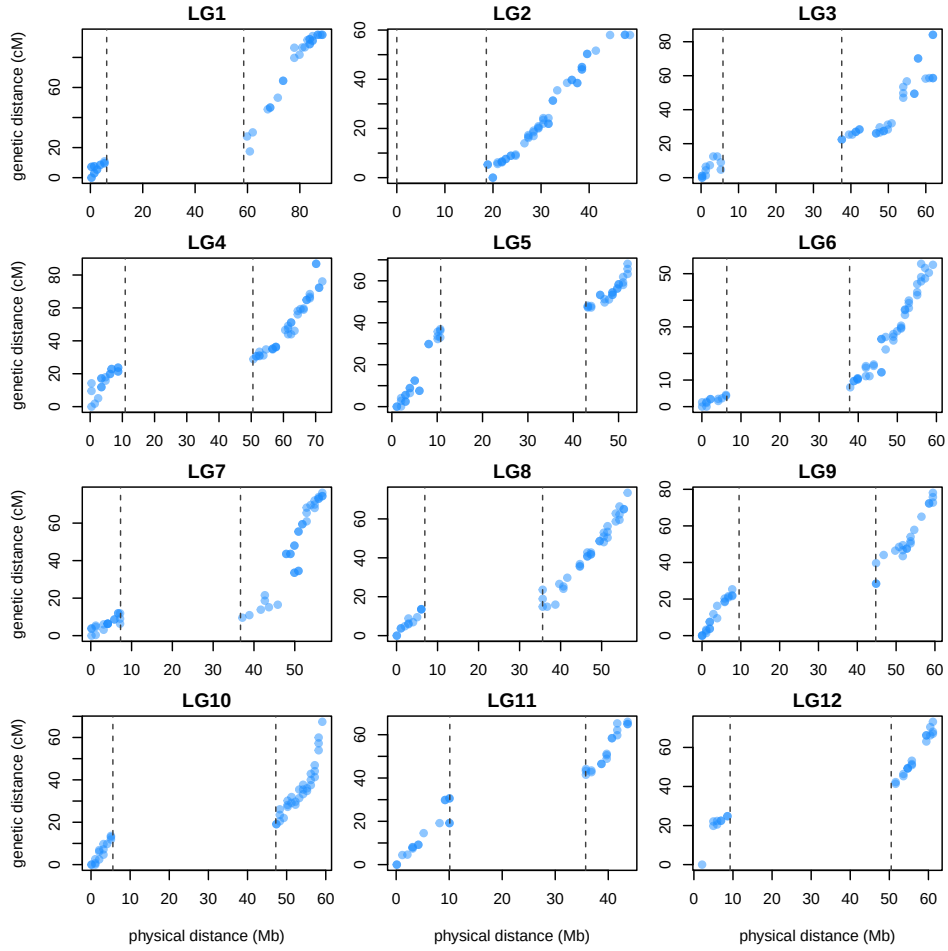


Figure 9. Plot of the genetic location (cM) vs. physical position (Mb) of PotatoMASH loci across the chromosomes. The vertical dotted lines represent the centromeric regions of the chromosomes.

4. Discussion

4.1. 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 [9] posited that, based on adopting an LD decay threshold of r^2 = 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. [15] to deal with this, using SMAP haplotype-sites [18] 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 [45]. 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.

4.2. 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 [46] 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.

4.3. 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. [8], 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 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 [8], 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).

4.4. 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 [2] 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.

4.5. 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. How-

ever, 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 [4]. 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 [47]. 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 [48], and separately, it has also been shown that allele dosage information can improve predictive abilities in comparison to using diploidized markers in polyploids [49]. 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 [50].

5. 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.

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

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.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 797162. This work was also supported by the DIFFUGAT project funded under EU the Horizon 2020 ERA-NET Cofund SusCrop (Grant No. 771134), being part of the Joint Programming Initiative on Agriculture, Food Security and Climate Change (FACCE-JPI), and funding from the Department of Agriculture, Food and the Marine (DAFM grant 2017/EN/109).

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.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

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
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

Table A1. *Cont.*

PotatoMASH Step and Supplies	Provider/Code	Pack Price (EUR)	Pack Units	Preps	Sample Cost (EUR)
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: *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.*					0.125
Total cost per sample:					**4.882**
Without library normalization:					3.759

References

1. Hamilton, J.P.; Hansey, C.N.; Whitty, B.R.; Stoffel, K.; Massa, A.N.; Van Deynze, A.; De Jong, W.S.; Douches, D.S.; Buell, C.R. Single nucleotide polymorphism discovery in elite North American potato germplasm. *BMC Genom.* **2011**, *12*, 302. [CrossRef] [PubMed]
2. Vos, P.G.; Uitdewilligen, J.G.; Voorrips, R.E.; Visser, R.G.; van Eck, H.J. Development and analysis of a 20 K SNP array for potato (*Solanum tuberosum*): An insight into the breeding history. *Theor. Appl. Genet.* **2015**, *128*, 2387–2401. [CrossRef] [PubMed]
3. Sverrisdóttir, E.; Byrne, S.; Sundmark, E.H.R.; Johnsen, H.Ø.; Kirk, H.G.; Asp, T.; Janss, L.; Nielsen, K.L. Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. *Theor. Appl. Genet.* **2017**, *130*, 2091–2108. [CrossRef]
4. Byrne, S.; Meade, F.; Mesiti, F.; Griffin, D.; Kennedy, C.; Milbourne, D. Genome-wide association and genomic prediction for fry color in potato. *Agronomy* **2020**, *10*, 90. [CrossRef]
5. Wickland, D.P.; Battu, G.; Hudson, K.A.; Diers, B.W.; Hudson, M.E. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinform.* **2017**, *18*, 586. [CrossRef] [PubMed]
6. Uitdewilligen, J.G.; Wolters, A.M.A.; D'hoop, B.B.; Borm, T.J.; Visser, R.G.; Van Eck, H.J. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE* **2013**, *8*, e62355. [CrossRef]
7. Jupe, F.; Witek, K.; Verweij, W.; Śliwka, J.; Pritchard, L.; Etherington, G.J.; Maclean, D.; Cock, P.J.; Leggett, R.M.; Bryan, G.J.; et al. Resistance gene enrichment sequencing (R en S eq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J.* **2013**, *76*, 530–544. [CrossRef]
8. Campbell, N.R.; Harmon, S.A.; Narum, S.R. Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Resour.* **2015**, *15*, 855–867. [CrossRef]
9. Vos, P.G.; Paulo, M.J.; Voorrips, R.E.; Visser, R.G.; van Eck, H.J.; van Eeuwijk, F.A. Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* **2017**, *130*, 123–135. [CrossRef]
10. Sharma, S.K.; MacKenzie, K.; McLean, K.; Dale, F.; Daniels, S.; Bryan, G.J. Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3 Genes Genomes Genet.* **2018**, *8*, 3185–3202. [CrossRef]
11. Consortium, P.G.S. Genome sequence and analysis of the tuber crop potato. *Nature* **2011**, *475*, 189.
12. Wolters, A.M.A.; Uitdewilligen, J.G.; Kloosterman, B.A.; Hutten, R.C.; Visser, R.G.; van Eck, H.J. Identification of alleles of carotenoid pathway genes important for zeaxanthin accumulation in potato tubers. *Plant Mol. Biol.* **2010**, *73*, 659–671. [CrossRef] [PubMed]
13. Uitdewilligen, J. *Discovery and Genotyping of Existing and Induced DNA Sequence Variation in Potato*; Wageningen University and Research: Wageningen, The Netherlands, 2012.
14. Uitdewilligen, J.; Wolters, A.; van Eck, H.; Visser, R. Allelic variation for alpha-Glucan Water Dikinase is associated with starch phosphate content in tetraploid potato. *Plant Mol. Biol.* **2022**, *108*, 1–12. [CrossRef]
15. Tinker, N.A.; Bekele, W.A.; Hattori, J. Haplotag: Software for haplotype-based genotyping-by-sequencing analysis. *G3 Genes Genomes Genet.* **2016**, *6*, 857–863. [CrossRef] [PubMed]
16. Baral, K.; Coulman, B.; Biligetu, B.; Fu, Y.B. Advancing crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] breeding through genotyping-by-sequencing and genomic selection. *PLoS ONE* **2020**, *15*, e0239609. [CrossRef]
17. Canales, F.J.; Montilla-Bascón, G.; Bekele, W.; Howarth, C.; Langdon, T.; Rispail, N.; Tinker, N.A.; Prats, E. Population genomics of Mediterranean oat (*A. sativa*) reveals high genetic diversity and three loci for heading date. *Theor. Appl. Genet.* **2021**, *134*, 2063–2077. [CrossRef]

18. Schaumont, D.; Veeckman, E.; Van der Jeugt, F.; Haegeman, A.; van Glabeke, S.; Bawin, Y.; Lukasiewicz, J.; Blugeon, S.; Barre, P.; de la O Leyva-Perez, M.; et al. Stack Mapping Anchor Points (SMAP): A versatile suite of tools for read-backed haplotyping. *bioRxiv* **2022**. [CrossRef]
19. Tang, X.; de Boer, J.M.; van Eck, H.J.; Bachem, C.; Visser, R.G.; de Jong, H. Assignment of genetic linkage maps to diploid Solanum tuberosum pachytene chromosomes by BAC-FISH technology. *Chromosome Res.* **2009**, *17*, 899–915. [CrossRef]
20. Sharma, S.K.; Bolser, D.; de Boer, J.; Sønderkær, M.; Amoros, W.; Carboni, M.F.; D'Ambrosio, J.M.; de la Cruz, G.; Di Genova, A.; Douches, D.S.; et al. Construction of reference chromosome-scale pseudomolecules for potato: Integrating the potato genome with genetic and physical maps. *G3 Genes Genomes Genet.* **2013**, *3*, 2031–2047. [CrossRef]
21. Meade, F.; Byrne, S.; Griffin, D.; Kennedy, C.; Mesiti, F.; Milbourne, D. Rapid development of KASP markers for disease resistance genes using pooled whole-genome resequencing. *Potato Res.* **2020**, *63*, 57–73. [CrossRef]
22. Hardigan, M.A.; Crisovan, E.; Hamilton, J.P.; Kim, J.; Laimbeer, P.; Leisner, C.P.; Manrique-Carpintero, N.C.; Newton, L.; Pham, G.M.; Vaillancourt, B.; et al. Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell* **2016**, *28*, 388–405. [CrossRef]
23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
24. Kofler, R.; Orozco-terWengel, P.; De Maio, N.; Pandey, R.V.; Nolte, V.; Futschik, A.; Kosiol, C.; Schlötterer, C. PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE* **2011**, *6*, e15925. [CrossRef] [PubMed]
25. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. *Nat. Biotechnol.* **2011**, *29*, 24–26. [CrossRef] [PubMed]
26. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
27. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3—New capabilities and interfaces. *Nucleic Acids Res.* **2012**, *40*, e115–e115. [CrossRef]
28. He, C.; Holme, J.; Anthony, J. SNP genotyping: The KASP assay. In *Crop Breeding*; Humana Press: New York, NY, USA, 2014; pp. 75–86.
29. Yuan, J.; Bizimungu, B.; De Koeyer, D.; Rosyara, U.; Wen, Z.; Lagüe, M. Genome-wide association study of resistance to potato common scab. *Potato Res.* **2020**, *63*, 253–266. [CrossRef]
30. Brown, S.S.; Chen, Y.W.; Wang, M.; Clipson, A.; Ochoa, E.; Du, M.Q. PrimerPooler: Automated primer pooling to prepare library for targeted sequencing. *Biol. Methods Protoc.* **2017**, *2*, bpx006. [CrossRef]
31. Magoč, T.; Salzberg, S.L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **2011**, *27*, 2957–2963. [CrossRef]
32. Gordon, A.; Hannon, G. Fastx-Toolkit. FASTQ/A Short-Reads Pre-Processing Tools. 2010. Available online: http://hannonlab.cshl.edu/fastx_toolkit (accessed on 26 August 2022).
33. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987–2993. [CrossRef]
34. Rosyara, U.R.; De Jong, W.S.; Douches, D.S.; Endelman, J.B. Software for genome-wide association studies in autopolyploids and its application to potato. *Plant Genome* **2016**, *9*, 1–10. [CrossRef] [PubMed]
35. Zhou, Q.; Tang, D.; Huang, W.; Yang, Z.; Zhang, Y.; Hamilton, J.P.; Visser, R.G.; Bachem, C.W.; Buell, C.R.; Zhang, Z.; et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **2020**, *52*, 1018–1023. [CrossRef]
36. Meade, F.; Hutten, R.; Wagener, S.; Prigge, V.; Dalton, E.; Kirk, H.G.; Griffin, D.; Milbourne, D. Detection of novel QTLs for late blight resistance derived from the wild potato species *Solanum microdontum* and *Solanum pampasense*. *Genes* **2020**, *11*, 732. [CrossRef]
37. Bourke, P.M.; van Geest, G.; Voorrips, R.E.; Jansen, J.; Kranenburg, T.; Shahin, A.; Visser, R.G.; Arens, P.; Smulders, M.J.; Maliepaard, C. polymapR—Linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics* **2018**, *34*, 3496–3502. [CrossRef] [PubMed]
38. Preedy, K.; Hackett, C. A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling. *Theor. Appl. Genet.* **2016**, *129*, 2117–2132. [CrossRef] [PubMed]
39. Zheng, C.; Amadeu, R.R.; Munoz, P.R.; Endelman, J.B. Haplotype Reconstruction in Connected Tetraploid F1 Populations. *Genetics* **2021**, *219*, iyab106. doi: 10.1093/genetics/iyab106. [CrossRef] [PubMed]
40. Rouppe Van der Voort, J.; Van der Vossen, E.; Bakker, E.; Overmars, H.; Van Zandvoort, P.; Hutten, R.; Klein Lankhorst, R.; Bakker, J. Two additive QTLs conferring broad-spectrum resistance in potato to *Globodera pallida* are localized on resistance gene clusters. *Theor. Appl. Genet.* **2000**, *101*, 1122–1130. [CrossRef]
41. van Eck, H.J.; Vos, P.G.; Valkonen, J.P.; Uitdewilligen, J.G.; Lensing, H.; de Vetten, N.; Visser, R.G. Graphical genotyping as a method to map Ny (o, n) sto and Gpa5 using a reference panel of tetraploid potato cultivars. *Theor. Appl. Genet.* **2017**, *130*, 515–528. [CrossRef]
42. Prodhomme, C.; Vos, P.G.; Paulo, M.J.; Tammes, J.E.; Visser, R.G.; Vossen, J.H.; van Eck, H.J. Distribution of P1 (D1) wart disease resistance in potato germplasm and GWAS identification of haplotype-specific SNP markers. *Theor. Appl. Genet.* **2020**, *133*, 1859–1871. [CrossRef]

43. Grech-Baran, M.; Witek, K.; Szajko, K.; Witek, A.I.; Morgiewicz, K.; Wasilewicz-Flis, I.; Jakuczun, H.; Marczewski, W.; Jones, J.D.; Hennig, J. Extreme resistance to Potato virus Y in potato carrying the Rysto gene is mediated by a TIR-NLR immune receptor. *Plant Biotechnol. J.* **2020**, *18*, 655–667. [CrossRef]
44. Tang, D.; Jia, Y.; Zhang, J.; Li, H.; Cheng, L.; Wang, P.; Bao, Z.; Liu, Z.; Feng, S.; Zhu, X.; et al. Genome evolution and diversity of wild and cultivated potatoes. *Nature* **2022**, *606*, 535–541. [CrossRef] [PubMed]
45. Willemsen, J. The Identification of Allelic Variation in Potato. Ph.D. Thesis, Wageningen University and Research, Wageningen, The Netherlands, 2018.
46. Bao, Z.; Li, C.; Li, G.; Wang, P.; Peng, Z.; Cheng, L.; Li, H.; Zhang, Z.; Li, Y.; Huang, W.; et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. *Mol. Plant* **2022**, *15*, 1211–1226. [CrossRef] [PubMed]
47. Selga, C.; Koc, A.; Chawade, A.; Ortiz, R. A bioinformatics pipeline to identify a subset of SNPs for genomics-assisted potato breeding. *Plants* **2021**, *10*, 30. [CrossRef] [PubMed]
48. Sallam, A.H.; Conley, E.; Prakapenka, D.; Da, Y.; Anderson, J.A. Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. *G3 Genes Genomes Genet.* **2020**, *10*, 2265–2273. [CrossRef]
49. Batista, L.G.; Mello, V.H.; Souza, A.P.; Margarido, G.R. Genomic prediction with allele dosage information in highly polyploid species. *Theor. Appl. Genet.* **2021**, *135*, 723–739. [CrossRef]
50. Slater, A.T.; Cogan, N.O.; Hayes, B.J.; Schultz, L.; Dale, M.F.B.; Bryan, G.J.; Forster, J.W. Improving breeding efficiency in potato using molecular and quantitative genetics. *Theor. Appl. Genet.* **2014**, *127*, 2279–2292. [CrossRef]