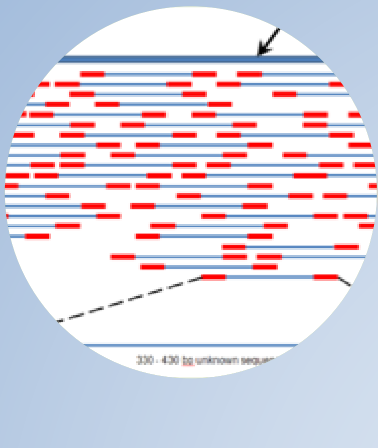
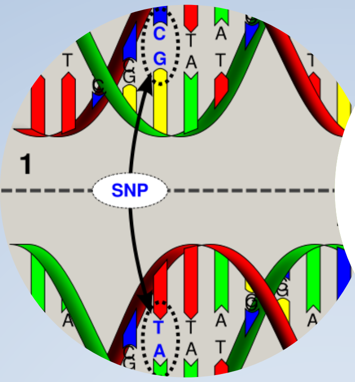
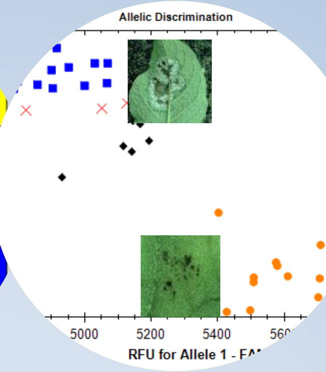






Potato late blight resistance gene, *Rpi-cap1*: haplotype-specific SNPs mining and validation on segregating population

Mainali, Ram

Wageningen University and Research
2018

Potato late blight resistance gene, *Rpi-cap1:* haplotype-specific SNPs mining and validation on segregating population

Ram P Mainali

881212540130

Supervisor:

Dr. Jack Vossen

Examinors:

Dr. Jack Vossen

Dr. Henk Schouten

Course: MSc Thesis Plant Breeding (PBR-80436)

for partial fulfillment of Master of Sciences in Plant Sciences Specialization Plant Breeding and Genetic Resources

Breeding for resistance in Solanaceae research group

Laboratory of Plant Breeding

Wageningen University and Research

January 2018

Front cover: Figurative view of paired-end sequencing, example of SNPs, clustering of allele during KASP genotyping, healthy potato (from left to right)

Back cover: Saying about farmers, scientists, and agriculturists (collected from internet and modified)

Acknowledgements

I started current thesis work with my personal objective of gaining insight on bio-informatics and link it to molecular breeding work. Before, I had a somehow little bit of theoretical concept with hardly any experience with bio-informatics stuff. The current thesis project is the example how motivation, support, and a good working group made the work possible. Of course, my first sincere gratitude goes to my supervisor Dr. Jack Vossen for his help, guidance, and feedback throughout the entire project. You are my best teacher who makes the environment for self-learning and critical thinking. I am indebted to Charlotte Prodhomme for her whole-hearted support and feedback to start my journey in Linux-environment. I would like to extend my heartfelt gratitude to Dr. Danny Esselink whom I met mid-way during my journey. I would like to thank Dr. Yuling Bai, my study advisor for her support during entire Wageningen living and study. Many thanks go to my examiner, Dr. Henk Schouten for his kind help and feeback. I would like to extend my acknowledgments to Isolde Bertram, Dirk Jan Huigen, Chen Xizheng, Gert van Arkel, Wageningen University and Research, Nepal Agricultural Research Council (NARC), Nepal, my family, Laboratory of Plant Breeding and Genetics, Netherland Fellowship Programs/Nuffic.

I will miss cappuccino, crowded breeder’s hall, and our motivative team.

Hope you enjoy my short journey on how bio-informatics and molecular breeding goes together.

Abbreviations

∩	Intersection
∪	Union
Avr	Avirulence
BAM	Binary alignment/map
BR	Resistant bulk
BSA	Bulk segregant analysis
BSKA	Bulked segregant k-mers analysis
BWA	Burrows-wheeler aligner
CAPS	Cleaved amplified polymorphic sequence
DM	Double monoploid
HR	Hypersensitive response
InDel	Insertion/delection
KASP	Kompetitive allele specific PCR
NB-NLR	Nucleotide-binding site leucine-rich repeat
PR	Resistant parent
QTL	Quantitative trait locus
R gene	Resistance gene
R haplotype	Resistance haplotype
R1	Forward read
R2	Reverse read
S haplotype	Susceptible haplotype
SAM	Sequence alignment/map
SNR	Signal-to-noise ratio

Abstract

Phytophthora infestans of oomycete group is considered as one of the most devastating pathogens that cause huge economic losses in agriculture by causing late blight on tomato and potato. Introgression of dominant *R* gene from wild germplasms through marker-assisted introgression breeding has been widely used to tackle this pathogen. DNA bulks of resistant (BR 7358) and susceptible segregating offspring plants (BS 7358) were derived from an intraspecific cross between *S. capsicibaccatum* resistant (Cap536-1) and *S. capsicibaccatum* susceptible (Cap564-3) plants. In the present study, next generation sequencing (NGS) based bulked segregant analysis (BSA) was employed using NGS data on whole genome sequencing of these bulks and the resistant parent to identify and validate the haplotype-specific SNPs associated to a dominant *Rpi-cap1* gene from *Solanum capsibaccatum*. We developed five different bulked segregant k-mers analysis (BSKA) to select resistance or susceptible bulk-specific k-mers from the causal genomic region. The 0-2 Mb interval on chromosome 11 and 20800-20950 Kb interval on chromosome 0 was determined as candidate gene region. We identified the best BSKA approach and single copy k-mers and reads from resistance and susceptible haplotype for *de-novo* assembly and haplotype-based variant detection. We determined 3261 and 4432 unique SNPs in resistance-specific haplotype and susceptible-specific haplotype relative to reference genome DM, respectively. Among 2315 susceptible haplotype-specific SNPs relative to *de-novo* assembled R contigs, only the SNPs present in those contigs that are anchored to a region of interest in reference genome DM was used to design KASP primer sets. Twelve KASP primer sets which are polymorphic to resistance trait were developed, however, only three were validated as informative markers to resistance trait in a F1 population in which *Rpi-cap1* is segregating. Validated markers, linked to *Rpi-cap1* were also tested on Athlete x Queen Anne population. This analysis showed that the late blight resistance and Avrcap1 response from Athlete were probably not caused by the *Rpi-cap1* gene. The present study found that the BSKA combined with *de-novo* assembly and haplotype-based variant calling is an efficient technique to determine the region of gene and SNPs associated with disease resistance. Future research could start with other potential markers and variant results that were suggested in the present study to fine map the *Rpi-cap1* gene. We recommend genetic mapping followed by map-based cloning of *Rpi-cap1* gene for gene pyramiding or temporal and spatial rotation of *R* genes for durable late blight management.

Keywords: Bulked segregant analysis (BSA), *de-novo*-assembly, haplotype, KASP genotyping, *Rpi-cap1*, variant calling

Table of Contents

List of Table

List of Figure

List of appendices

1. INTRODUCTION

1.1 Background

Potato (*Solanum tuberosum* L.) is a major staple food crop worldwide after corn, paddy, and wheat. Europe and Asia are the foremost potato producing regions and contributed over 80% of global production (FAOSTAT, 2016). The cultivation and production of potato is growing in many developing countries, mainly in Asia, underpinning the imperative role of potato to meet the needs of growing human population (Birch *et al.*, 2012; FAOSTAT, 2016). Nonetheless, potato production is threatened by a wide array of biotic and abiotic factors. Biotic factors such as microbial diseases and insect pests are becoming more problematic especially due to climate change (Birch *et al.*, 2012). Besides, cropping patterns, change in agronomic practices, for example, monoculture favors expansion of biotic agents (Fiers *et al.*, 2012). In this regard, understanding of host-pathogen interaction is the key to unravel the scientific ground for sustainable disease management.

Pathogenicity of the pathogen and the corresponding host response are pivotal to determine the degree of host resistance and the level of avirulence or virulence of the pathogens during the host-pathogen interaction (Flor, 1971). The gene for gene hypothesis proposed by Flor in the 1940s is the most archetypal and well-studied model that describes host-pathogen interaction in the process of evolution. It states that for each major resistance (*R*) gene providing resistance in host plant have corresponding avirulence gene in the pathogen (Flor 1971) which contributes to pathogenicity (Staskawicz *et al.*, 1995). Later, it was illustrated that the effectors produced by plant pathogens are the driving force for plant-microbe interaction (Hogenhout *et al.*, 2009). The resistance protein of the plant recognizes a particular effector protein, which in turn activates the plant defence and stops the pathogen growth, often culminating in a (visual) hypersensitive response (HR) (Jones and Dangl, 2006). Most of the *R* genes in plants encode nucleotide-binding site leucine-rich repeat (NLR) protein (Lozano *et al.*, 2012). This NLR protein recognizes associated effector protein secreted by a diverse group of pathogens such as nematodes, oomycetes, fungi, bacteria, viruses, and insects (McHale *et al.*, 2006). In general, these *R* genes reside in complex clusters of similar sequences, which act as a pool providing variation for resistance specificities. It consists of wide ranges of hypervariable potential ligand-binding sites, which generate and maintain resistance specificities against ever-changing pathogens largely by interallelic recombination and for instances by unequal crossing-over (Michelmore and Meyers, 1998).

Potato late blight is considered one of the most important devastating diseases in potato, which is caused by one of the oomycete pathogens, *Phytophthora infestans* (Mont.) de Bary. The intensity of devastation has been chronologically recorded since the identification of the *P. infestans* associated with potato late blight causing Irish famine to this date. *P. infestans* is heterothallic in nature. Two dominant mating type strains type A1 and type A2 are prevalent in Europe and the most of the world (Drenth *et al.*, 1993, Hwang *et al.*, 2014). The sexual recombination between these two mating types results in variable oospores, providing tremendous complex genetic diversity thereby accelerating adaptation of pathogen genotypes with changing environment and farming condition (Drenth *et al.*, 1993; Li *et al.*, 2012a; Hwang *et al.*, 2014). These oospores (inoculums) governing sexual cycle can continue the pathogens' survival between two growing seasons. Moreover, the infected plant is the source of asexual sporangia that produce motile zoospores (Figure 1) (Govers *et al.*, 1997). Sporangia and zoospores are the sources for tuber or leaf infection and subsequent disease development. The increasing inoculum level overtime and complexities inflict sizable damage on all parts of the plants such as tuber, stem, and leaf and may impart huge damage on crop yield (Govers *et al.*, 1997; Tsedaley, 2014).

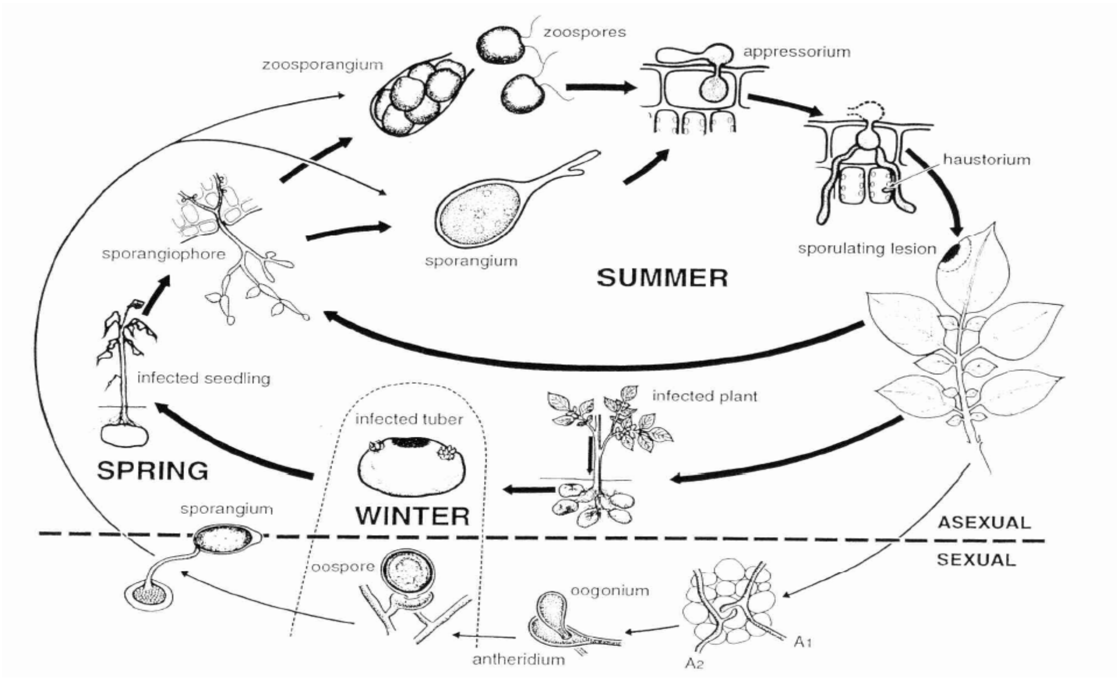


Figure 1 Potato late blight disease cycle (source: Govers *et al.*, 1997)

Management of *P. infestans* is still challenging across the globe. The use of fungicides has been the most dominant control measure since the mid-20th century. However, there is an increasing issue of the use of fungicides on human and microbial health, environment, and farming budget (Kromann *et al.*, 2011). Still, the increasing resistance of pathogens against systematic fungicides combined with the legal restrictions on their use throughout the globe has urged growers to seek alternatives for pesticides (Potts 1990; Tamm *et al.,* 2004; Hwang *et al.*, 2014). Other sustainable approaches to management, for example, site selection, mixed cropping, crop rotation may be the potential solution to late blight, however, always potato growth without chemical protection remains associated with production risks (De Buck *et al.*, 2001). In this regard, developing late blight resistant cultivars through resistance breeding could be one of the potent alternative management strategies, however, time, budget and associated risks (breakdown of resistance) should be considered during the design of breeding. Resistance breeding strategies rely on the introgression of *R* genes from wild *Solanum* germplasm or through marker-free cisgenesis. In general, the products of *R* genes trigger a hypersensitive response to cognate effectors (Avr) released by the pathogen. Upon Avr recognition, the growth and development of pathogen is restricted (Vleeshouwers *et al.*, 2011). However, the durability of these introgressed *R* genes in host plant is a question as pathogen effectors can easily escape from recognition from host *R* gene and therefore resistant cultivars did not hold stable resistance on the long run. This situation is more imminent when breeding strategy relies on a single dominant *R* gene as this will exert selective pressure on pathogens to lose Avr, causing a quick break down of the resistance (McDonald and Linde, 2002; Fry, 2008). This situation prompted scientists to explore genes showing broad-spectrum quantitative or partial resistance (van der Vossen *et al.,* 2005). Besides, breeders are also looking for a diversity of *R* genes for multiple *R* gene staking or temporal and spatial rotation of these *R* genes.

Several *R* genes or their combinations were deployed to develop resistant commercial potato cultivars to combat *P. infestans* in the past and this technique is still in practice. At least 11 *R* genes from *Solanum demissum* have been introduced in commercial potato cultivars (Gebhardt and Valkonen, 2001). The new generation of late blight *R* genes found in other germplasms besides *S. demissum* has become one of the important targets for the breeders since last decade. The new generation *R* genes have been found in different wild relatives of potato, for example, *Rpi-blb1*, *Rpi-blb2, and Rpi-blb3* from *Solanum bulbocastanum* (van der Vossen *et al.*, 2003; van der Vossen *et al.*, 2005), *Rpi-ber* from *Solanum* *berthaultii* Hawkes. (Rauscher *et al.*, 2006), *Rpi-mcd1* from *Solanum microdontum* Bitter. (Tan *et al.*, 2008), *Rpi-cap1* from *Solanum capsicibaccatum* Card. (= *S. circaeifolium*) etc. This diversity of new generation of late blight *R* genes increase the probability of finding best durable resistance in potato through staking best compatible combination of multiple genes in current breeding programs. *Rpi-cap1* gene derived from *S capsicibaccatum* Card. (2n=2x=24) is the topic of this study.

A novel resistance gene from wild diploid germplasm, *S. capsicibaccatum*

Wild Bolivian diploid species, *Solanum capsicibaccatum* Card. contain two subspecies such as *circaeifolium* and *quimense* (Hawkes and Hjerting 1989). These species contain new generation *Rpi* gene *i.e. Rpi-cap1*. This gene confers monogenic resistance, but also the high level of broad-spectrum resistance characteristics (Verzaux *et al.*, 2012). *Rpi-cap1* is located on the long arm of the chromosome 11 in a cluster of *N*-like sequences (Jacobs *et al.*, 2010; Verzaux *et al.*, 2012) (Figure 2). Therefore, unlike all other major genes, which belongs to the CC-NBS-LRR (CNL) class (Hein *et al.*, 2009), it is hypothesized that the *Rpi-cap1* is belongs to the class TIR-NBS-LRR (TNL) (Verzaux *et al.*, 2010). Intriguingly, the distal end of the chromosome 11 was termed as a hotspot as they harbour *R* genes or QTL for viruses, nematodes and disease pest including *P. infestans* (Gebhardt and Valkonen, 2001).

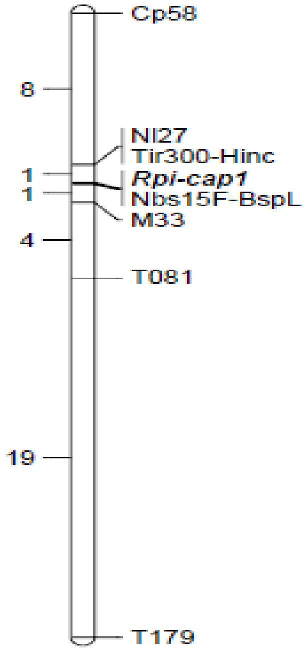


Figure 2 The genetic map of new generation *R* gene, *Rpi-cap1* on chromosome 11. The number on left side of the map shows the number of recombinants out of 900 progeny population. Used maker type in the study was CAPS except Nbs15F-BspL and Tir300-Hinc (Source: Verzaux *et al.*, 2012)

BSA approach and *de-novo* assembly

With the advent of producing high throughput Next Generation Sequencing (NGS) data on the whole genome, the gene mapping is becoming more efficient in terms of cost and time. Bulked segregant analysis (BSA) is a rapid and convenient method to locate candidate gene or QTL with chromosome position associated with a trait specific phenotype, for example, mutant or disease resistance (Giovannoni *et al.*, 1991; Michelmore *et al.*, 1991). Due to the high resolution of parallel sequencing techniques, this method is becoming popular to identify polymorphic loci as well as their allele frequency (Magwene *et al.*, 2011). Instead of a conventional study of all entities from the sample population (conventional mapping study), BSA approach examines sequencing data on selected pooled entities showing extreme phenotype and therefore allow to derive target region from the whole genome that most likely harbour resistance gene. The target regions showing genetic variation are useful for gene mapping (Quarrie *et al.*, 1999; Zou *et al.*, 2016). The proper BSA design consists development of large segregating population, using large sample size, representing each phenotypic pool, and finally selecting DNA markers that have higher coverage and specific to target trait (Takagi *et al.*, 2013; Zou *et al.*, 2016). This approach is useful in genetics, genomics, and other plant breeding activities (Zou *et al.*, 2016). During the process, the mapping of the target reads derived from BSA approach relies on assembling similar NGS reads against the reference genome, however, the reference genome is not always available or may not be complete or some time may not be handy to use. In this regard, construction of haplotype sequence using NGS data would be promising.

De-novo assembled NGS reads create an original draft sequence which was unknown before. In general, the short read assemblers use either of two major classes of Lander-Waterman model based assembly algorithms such as overlap-layout-consensus (OLC) and de Bruijn graphs (DBH) based algorithms (Li *et al.*, 2012b). OLC first found overlaps (O) among all the short reads followed by their layout (L) and finally, the consensus (C) sequence structure in the form of contigs is produced. De Bruijn graph is based on alignments of k-mers which are shorter than read length and created contigs (Schatz *et al.*, 2009; Li *et al.*, 2012b). Primarily, both algorithms processing efficiency and the result depend on sequencing depth and read length. Genome characteristics, like heterozygosity, large copy numbers, transposon and other repetitive sequences impair proper assembly of short reads resulting inaccurate *de-novo* assembly result (Jupe *et al.*, 2013).

Kompetitive Allele-Specific PCR (KASP) genotyping

Following the paradigm shift in Next Generation Sequencing and development of bioinformatics tools for metadata analysis and management, the determination of the genetic variations is becoming more efficient. The genetic variation, for example, SNPs, InDels between any group of genotypes or haplotypes or bulks potentially allow development of markers specific to target trait or gene and may replace the old type markers, like SSR marker. However, depending upon the objectives, one may consider the ease of use, cost-effectiveness, lower error rate, performance, throughput, flexibility, assay capability and requirements during marker design (Semagn *et al.*, 2013). It is apparent that genotyping by next-generation sequencing is one of the growing methods for variant genotyping (Ertiro *et al.*, 2015).

Kompetitive Allele-Specific PCR (KASP^{TM}) genotyping is a competitive allele-specific polymerase chain reaction (PCR) based high throughput uniplex SNPs or InDel genotyping platform. This platform is a gel-free assessment in a closed tube (Neelam *et al.*, 2013) based on allele (SNP) specific oligo extension and successive fluorescence resonance energy transfer (FRET) (Kumpatla *et al.*, 2012; Semagn *et al.*, 2013). The KASP reaction uses one common reverse primer and two allele-specific forward primers or vice versa. It uses pre-selected SNPs or target polymorphism and reduces the rate of error, cost and time than multiplexed chip-based technology if assay was performed for small to a modest number of SNPs. This system can be operated in the basic molecular laboratory and is effective to use in QTL mapping, genetic mapping, germplasm genotyping, allele mining, MAS breeding etc. (Neelam *et al.*, 2013; Semagn *et al.*, 2013). KASP provides the flexible choice if the markers are readily mapped to the specific genomic region of selected crossing cultivars in which gene of interest is going to be introduced.

1.2 Research rationale and objectives

Most of the cultivated potato cultivars are autotetraploid (2n=4x=48) with a basic chromosome number of 12. Potato has a high number of crossable wild relatives; more than 200 *Solanum* species so far known. These wild relatives vary in ploidy level ranging from diploid (2n=2x=24) to hexaploid (2n=6x=72) (Watanabe, 2015). During the process of domestication, the current potato cultivar lost a huge genetic variation leading to narrow a genetic base which is a likely the reason for the current late blight epidemics in potato. Owing to this fact, wild genetic pools are always of interests for breeders to introduce new or improved traits into the modern cultivars (Bradshaw, 2009). Genes have been introduced either through recurrent backcrossing from various wild germplasms (Kim *et al.*, 2012) or through cisgenesis utilizing the cloned genes (Schouten and Jacobsen, 2008). The introgression breeding or cisgenesis both require molecular markers, which will eventually help for rapid genotyping with associated multiple advantages (Tiwari *et al.*, 2013).

There are already few molecular markers showing polymorphism and linked with *Rpi-cap1* gene. Jacobs *et al.* (2010) used NBS profiling technique to develop cleaved amplified polymorphic sequence (CAPS) markers, for example, the resistance associated co-segregating marker Cp58. Besides, Verzaux et al. (2012) used *R* gene cluster directed profiling approach for developing CAPS markers that are either closely linked or co-segregating with this novel *Rpi-cap1* gene. These markers were developed from diploid *Solanum capsicibaccatum* which are probably less suitable for tetraploid potato (Jack Vossen, Personal Communication). With the advent of modern technologies of genome sequencing, whole genome sequencing is being more common and is one of the best starting points to explore new potential markers. The intraspecific cross between *S. capsicibaccatum* resistant (Cap536-1) and *S. capsicibaccatum* susceptible (Cap564-3) plants (described by Verzaux *et al.,* 2012) was expanded and their progenies were evaluated using late blight infection. A 1:1 segregation ratio of late blight resistant and susceptible genotypes was found. Even, it showed co-segregation with *Avr-cap1* response suggesting the presence of a single dominant *R* gene, *Rpi-cap1*. Recently DNA bulks of 11 resistant (BR 7358) and 11 susceptible offspring plants (BS 7358) were made (Appendix I; Jack Vossen, unpublished results). DNA of these bulks and the resistant parent (PR) were sent for whole genome shot-gun sequencing on the Illumina HiSeq-X platform. Paired-end reads of 151 bp from BR, BS and, PR were obtained with more than 15X expected coverage depth for each sample.

The genetic variation between two contrasting phenotypes (traits) determines the trait-specific variant (Michelmore *et al.,* 1991). These genetic variations, for example, SNPs and InDels are abundantly found in the potato genome (Potato Genome Sequencing Consortium, 2011). These forms of genetic variation are highly amenable to current advanced genotyping platforms (genotyping by sequencing). In this regard, the first objective of the present study is **to identify SNPs specific to the resistance haplotype of wild diploid species, *S. capsicibacatum***. We used bulked segregant k-mer analysis (BSKA) approach combined with *de-novo* assembly and haplotype-based variant calling to mine true SNPs. We screened for SNPs that are present in resistance haplotype (R haplotype) specific reads but not present in susceptible haplotype (S haplotype) specific reads and other nine breeding germplasms mostly the commercial potato cultivars not containing *Rpi-cap1* gene.

Variant information can be potentially converted to markers. In this regard, marker development followed by their validation in segregating progeny is imperative. Hence, the second objective of the present study is **to design SNPs based markers and test them in parent and bulk members**. We developed Kompetitive Allele-Specific PCR (KASP^{TM}) marker utilizing unique SNPs associated with *de-novo* assembled contigs that anchored to a region of interest in reference genome DM. These markers were then validated in the individual members of BR and BS, but also the parent of cap7358 population followed by re-confirmation in additional offspring that were not in bulks.

Athlete is one of the few late blight resistant cultivars of potato, which is used for growing under organic condition. It is tetraploid potato cultivar developed by Agrico UK Ltd with parentage AR 99-263-5 x Miriam. Dr. Ronald Hutten from Wageningen University and Research made a cross between Athlete and Queen Anne and a 1:1 phenotypic segregation ratio was observed for resistance to late blight, which co-segregated with *Avr-cap1* responsiveness (Dr. Jack Vossen, unpublished data). It was expected that Athlete might have the same resistance gene, *Rpi-cap1* as *S. capsicibaccatum* accession which is most likely coming from interspecific hybrids during the breeding process. Alternatively, it may be possible that other *R* genes in Athlete recognize the same *Avr* protein. Hence, the third **objective of this present study is to determine whether the Athlete contains the same resistance gene (*Rpi-cap1*) as *S. capsibaccatum* accession?** For this purpose, we used validated KASP markers associated with resistance trait in a cap7358 population.

To address these objectives, the following research questions were formulated,

i. How to mine SNPs specific to resistance haplotype derived from intraspecific crosses of *S. capsibaccatum* using BSA approach?
ii. How to filter for informative SNPs from *S. capsibaccatum* to resistance haplotype?
iii. How to design haplotype-specific markers?
iv. Does Athlete contain the same *R* gene (*Rpi-cap1*) as resistant *S. capsicibacctum* accession?

2 METHODOLOGY

2.1 Data and server description

The Illumina NGS data on whole-genome sequencing of DNA from BR, BS, and PR were used for this study. Illumina paired-end short gun sequencing read of 2 * 151 bases for each bulk and PR were obtained from Jack's root directory. The computer algorithms bulked segregant analysis combined with *de-novo* assembly, and successive variant calling was done to mine SNPs using MobaXterm Personal Edition v6.6 (Unix utilities and X-server on Gnu/Cygwin) which is communicated remotely to the plant breeding server of Wageningen University and Research through SSH sessions. I kindly received the basic bash script from Charlotte Prodhomme and Dr. Danny Esselink. The software WinSCP was used to upload and download the file.

2.2 Bulk segregant k-mers analysis (BSKA) approaches using NGS data on whole genome Sequencing

2.2.1 Data processing

The quality of Illumina NGS raw reads of a parent and each bulk were checked using fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The fastQC resulted in summary tables and graphs, which were used to assess each data set. The multiQC was then used to compile the fastQC results into a single report (http://multiqc.info/). To get the perfect quality of reads, both forward sequencing reads (R1) and reverse sequencing reads (R2) were trimmed using read trimming tool called Trimmomatic-0.32. The low-quality bases at beginning and end of reads, Illumina adapters, read less than 70 bases were removed. Also, unpaired R1 and R2 reads were removed. All paired reads, passing QC, were used as input for BSKA algorithms (Figure 3).

We used software package GenomeTester4 for BSA approach. GenomeTester4 is written in the C programming language which runs from the command line on Linux or other Unix-like operating system. It consists of three programs: GlistMaker, GlistCompare, and GlistQuery. GlistMaker generates k-mers (short k-length sub-sequences from a DNA sequence reads) list from original sequencing reads. GlistCompare executes basic algebraic set operations such as union, intersection, complement/difference, etc. Using lists generated by GlistMaker and Glistcompare, third program GlistQuery searches for statistics or user-provided sequences (Kaplinski *et al.*, 2015).

The k-mers list of 31 nucleotides for each bulk and PR were derived from trimmed reads (R1 and R2). The k-mers list from both R1 and R2 were combined with GlistCompare (union function) program of GenomeTester4 toolkit package for each PR, BR, and BS. The read errors that were represented singly (error) were removed by the program Glistcompare (intersection function). The error-free k-mers that were represented at least two times (cut off or coverage threshold 2) were used to produce a k-mers histogram for each bulk and PR. The k-mers lists were then used to produce bulk or haplotype-specific reads through different bulk segregant k-mers analysis (BSKA) approaches. For each approach, the different coverage depth or cut-off were chosen to remove the least represented k-mers or keep the k-mers list above respective threshold, for example, haplotype with cut-off or coverage 6 keeps those unique k-mers that are repeated at least 6 times in the sequence dataset (exclude all the k-mers that represented 5 or less than 5 times).

2.2.2 Bulk specific k-mers mapping to reference genome to determine causal genomic region

To determine the causual genomic region associated with the trait of interest in present study, the list of k-mers retrieved from BSKA approach with above-mentioned coverage is mapped to the potato reference genome, *S. tuberosum* group Phureja DM1-3 516 R44 (PGSC v4.03 Pseudomolecules) (Potato Genome Sequencing Consortium, 2011) using Burrows-Wheeler Aligner (BWA) -backtrack which is short read alignment algorithm designed for mapping of Illumina NGS reads up to 100 bp (Li and Durbin, 2009). BWA yields new standard SAM (Sequence Alignment/Map), BAM (Binary Alignment/Map) format files which can be visualized in IGV. The mapping quality was checked using samtools by using BAM format file as input. The quality was double checked with qualimap_v2.2/qualimap-bamqc, which produced the summary of a qualimap report. The unique k-mers that mapped to the reference genome were counted per 1 Mb window size of potato genome (bed format of PGSC v4.03 Pseudomolecules) using BAM format file as input. The k-mers count per 1 Mb interval of all potato chromosomes was plotted in a graph using Microsoft Excel. The signal-to-noise ratio (SNR) was calculated through dividing average k-mers count of peak region (signal) by average k-mers count of the remaining region (noise) of the same chromosome. The detected region of interest was used for downstream *de-novo* assembly.

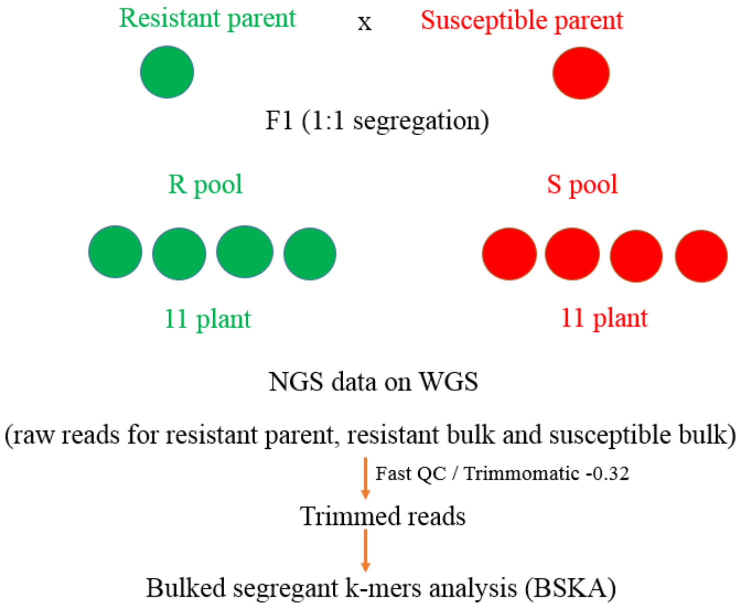

Resistant parent
x
Susceptible parent
F1 (1:1 segregation)
R pool
S pool
11 plant
11 plant
NGS data on WGS
(raw reads for resistant parent, resistant bulk and susceptible bulk)
Fast QC / Trimmomatic -0.32
Trimmed reads
Bulked segregant k-mers analysis (BSKA)

Figure 3 Research pipelines adopted before BSKA algorithms; the circle filled with green colour represents the resistant plant whereas circle filled with red colour represents the susceptible plant

2.3 De novo assembly and variant calling

2.3.1 Haplotype specific read retrieval and *de-novo* assembly

The k-mers that mapped to the reference genome in specific target region was retrieved for both R and S haplotype using samtools. The mapped k-mers to DM in BAM file was indexed and used as input for samtools. From k-mers lists, we selected only those k-mers having mapping quality score more than one and sorted in alphabetical orders. These k-mers were then used to retrieve the sequencing reads from associated bulk or PR that contained at least one k-mers from the list. For example, to retrieve the R haplotype-specific reads, first R1 and R2 reads in PR containing k-mers from our sorted list were retrieved. Similarly, R1 and R2 reads that contains at least one kmer were extracted from BR. Then we combined R1 of PR and BR using cat function. We also combined R2 form PR and BR using the same function. The same process was followed to retrieve the S haplotype-specific reads but the BS was used instead of BR. The combined R1 and R2 (in fastaq format) for each R haplotype and S haplotype were then used for *de-novo* assembly. The SOAP configuration file was prepared for both R and S haplotype reads. Average insert size was calculated by combined use of SOAPdenovo-V1.05, bwa-0.7.12/bwa mem, samtools and R studio (refer details at http://www.cbs.dtu. dk/courses/27626/Exercises/denovo _exercise.php). Finally, haplotype-specific *de-novo* assembly was done using the SOAPdenovo-V1.05

program. SOAPde-novo2 builds succinct De Bruijn graph using SOAPdenovo-127mer version or SOAPdenovo-63mer version. The first version supports k-mer size of ≤127 and consumes high memory, however, version SOAPdenovo-63mer supports k-mer size of ≤63 and significantly reduced the memory usage (Luo *et al.*, 2012; Luo *et al.*, 2015). We used k-mer sizes of 41, 51, 61, 71, and 81 for *de-novo* assembly. The assembly quality with these different k-mer sizes was evaluated using Assemblathon (Earl *et al.*, 2011). Based on assembly evaluation report, we determined the best assembly result for each R haplotype and S haplotype (appropriate k-value).

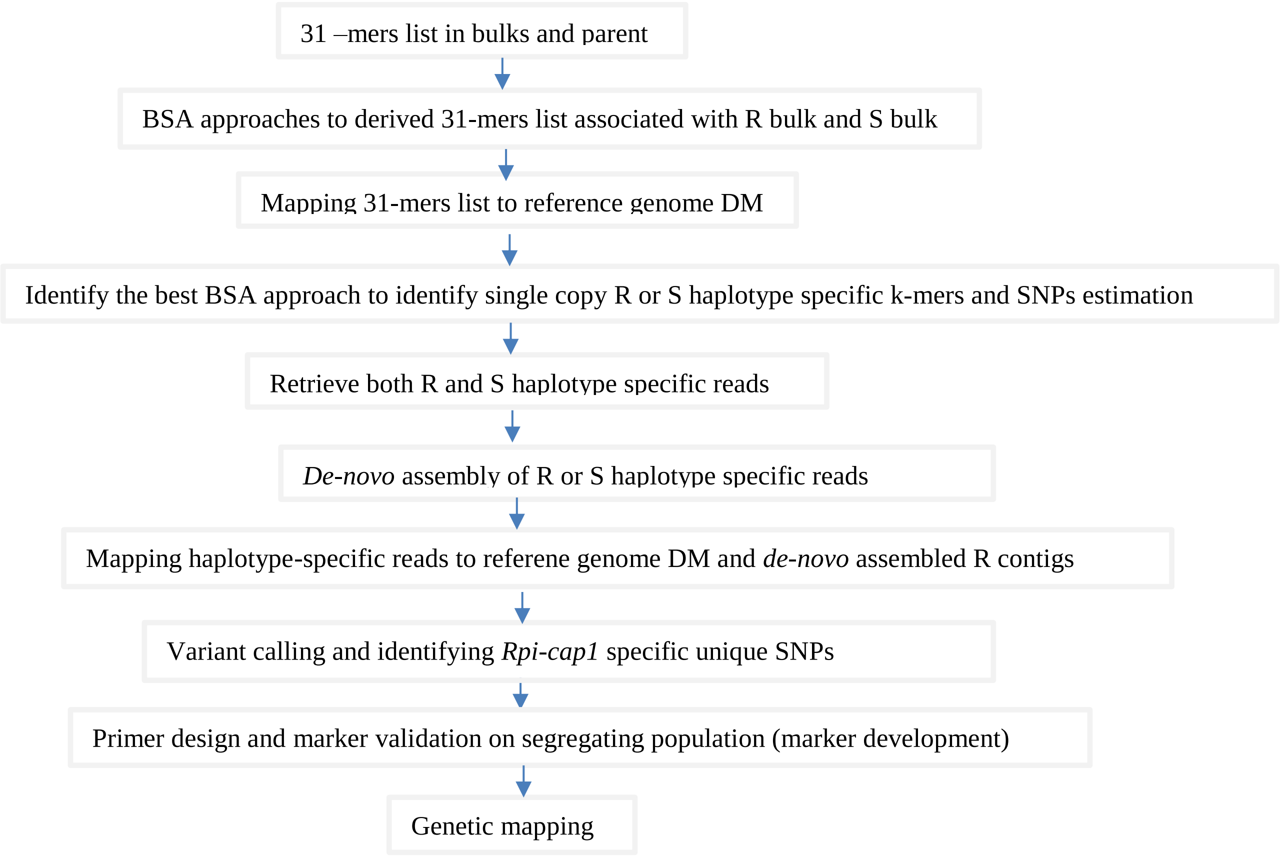

31 –mers list in bulks and parent
BSA approaches to derived 31-mers list associated with R bulk and S bulk
Mapping 31-mers list to reference genome DM
Identify the best BSA approach to identify single copy R or S haplotype specific k-mers and SNPs estimation
Retrieve both R and S haplotype specific reads
De-novo assembly of R or S haplotype specific reads
Mapping haplotype-specific reads to referene genome DM and *de-novo* assembled R contigs
Variant calling and identifying *Rpi-cap1* specific unique SNPs
Primer design and marker validation on segregating population (marker development)
Genetic mapping

Figure 4 Bulked segregant k-mers analysis, *de-novo* assembly and variant calling pipelines

2.3.2 Haplotype-based variant calling

The R haplotype-specific R1 and R2 reads which were produced and used in *de-novo* assembly were also used to read mapping against reference genome DM. Similarly, the S haplotype-specific trimmed R1 and R2 were mapped to DM. Nine different susceptible potato cultivars (Appendix II) were mapped to the DM reference genome after quality trimming. Alternatively, R1 and R2 from R haplotype-specific reads, S haplotype-specific reads and each nine susceptible cultivars specific reads were mapped separately against resistance specific contigs (*de-novo* R contigs). We used BWA-mem mapping program for above read mapping. It produced alignment file in BAM format which was used to visualize SNPs in IGV.

The alignment file in BAM format of R haplotype, S haplotype and other 9 susceptible cultivars in relative to reference DM and reference R contigs were detected using haplotype based variant detector tool, FreeBayes version v1.0.1-2-g0cb2697. This tool generates a variant report in VCF format using BAM alignment files and corresponding FASTA reference sequences as input (Garrison and Marth, 2012). To avoid false positive variant, filtering was done using different filtering criteria, for example, QUAL, depth (DP), observation count (AO) (Appendix III), however, less stringent filtering criteria was set for variant calling for 9 susceptible cultivars (due to lower allele frequency in tetraploid and different sequencing depth of NGS data among susceptible cultivars available for present study). The variant report was associated with filtering criteria tag from which we selected quality passed SNPs that are present only in R haplotype but not present in S haplotype and other 9 susceptible cultivars using Microsoft Office Excel, based on the relative position of the SNPs in DM. An outline of the variant calling process is depicted in Figure 4.

2.4 Primer Design

Two allele-specific forward (or reverse) primers and corresponding common reverse primer (or forward primer) were designed manually using Primer3Plus online software (http://www. bio inform atics.nl/cgi-bin/primer3plus/primer3plus.cgi). The primer standards such as 18-27 bp primer length, 50-200 bp PCR product length, 57-63 ^{0}C melting temperature (Tm), maximum 2 ^{0}C temperature difference, optimum 40-60% GC content, proper distribution of GC and AT-rich domains, absence of repeats, hairpin, intra-primer and inter-primer homology (complementary sequences) etc were provided as input. The primers were synthesized at Biolegio BV and stored at Laboratory of Plant Breeding and Genetics freezer 15 (left, third chamber from down).

2.5 Plant material

Following an intraspecific cross between diploid parental lines, *S. capsicibaccatum* resistant (Cap536-1) and *S. capsicibaccatum* susceptible (Cap564-3), the segregating F1 progeny (cap7358 population) were determined for resistance or susceptible phenotype against late blight pathogens and its effectors (1:1 segregation). There was a similar result for Athlete x Queen Anne tetraploid crossing. These plant materials of both diploid and tetraploid crossing were maintained *in-vitro* (Vossen, unpublished report). The tissue culture *in-vitro* plantlets of both cap7358 and Athlete x Queen Anne were kindly received from Isolde Bertram. These in-vitro plantlets of both populations were then transferred to a greenhouse with the help of Dirk Jan Huigen.

2.6 DNA extraction (Modified CTAB method)

DNA was extracted from parent and segregating progeny form cap7358 population and Athlete x Queen Anne population. Two very small young leafs were collected in 96 deep well block containing two Tungsten Carbide Bead (3 mm, Qiagen) and frozen in -80 ^{0}C. Deep well block was shaken for 2 minutes at 20 /s in the RETCH machine and again frozen in -80 ^{0}C at liquid nitrogen until DNA isolation. Freshly made 400 µl isolation buffer was added to deep well block and mixed manually. The block was then kept for 1 hour at 65 °C and mixed occasionally. The block was cooled to room temperature and 400 µl chloroform/isoamylalcohol (24:1) was added to each sample and mixed manually. The mixed samples were centrifuged at 6000 rpm for 20 minutes. The supernatant was transferred to new deep well block and 0.8 volume of isopropanol was added followed by mild mixing. Again, the new deep well block was centrifuged at 6000 rpm for 8 minutes followed by supernatant removal. 300 µl 70% ethanol was loaded to each sample and centrifuged for 5 min at 60000 rpm. The supernatant was removed and deep well block again centrifuged briefly. The remaining ethanol was removed by pipetting, and the pallet was dried at room temperature for a night. Finally, the pallet was dissolved in 75 µl MQ + RNase. The DNA was then transferred to small 96 well PCR plates and stored in refrigerator 15 at -20 ^{0}C (Supplementary file 1, File name: DNA storage from *capsibaccatum* project). 5 µl DNA was transferred to another 96 PCR plate and then diluted by 10x (45 ml MQ was added). 5 µl of 10x diluted DNA was loaded in 0.8% agarose gel well to determine the concentration of DNA through relative gel red staining intensity.

2.7 KASP assay

For PCR reactions, the DNA was diluted to 20ng/ µl in MQ using 10x diluted DNA electrophoresis report. The primer stock solution (100 µM/µl in MQ) was diluted 100x to obtain the final working concentration (1µM/µl). 9 µl of master mix was prepared by mixing each 1 µl of resistance allele-specific primer, susceptible allele-specific primer, common primer, MQ and 5 µl of KASP master reaction mix. Based on the number of DNA samples used for PCR, first, we calculated the amount of each reagent of master mix and mixed manually. The 9 µl master mix was loaded to 96 PCR plate followed by sample DNA loading (1 µl). Two non-template controls without DNA (MQ) were used as the check in each PCR run, by adding 1ul of MQ instead of DNA to the 9ul of master mix (Table 1). As a general procedure, during PCR, cycling conditions: 94 °C for 15 min, followed by 9 cycles of touchdown PCR: 94 °C for 20s, 61 °C for 1 minute with dropping rate 0.6 °C per cycle. Then, the 25 cycles of regular PCR: 94 °C for 20 s and 55°C for 1 minute was set followed by 37 °C for 1 minute. The fluorescence for each sample was then read by Bio-rad q-PCR machine. To determine the presence or absence of SNPs, the resistance specific allele was labeled with FAM using a 5'extension to the primer (gaaggtgaccaagttcatgct) whereas susceptible specific allele was labeled with HEX (5'extension: gaaggtcggagtcaacggatt). For confirmation of successful PCR, the PCR products were then allowed for electrophoresis in 2% agarose gels and visualized through gel red fluorescence.

Table 1 Reagents that are used in KASP genotyping

KASP genotyping reaction constitute		Doses for each sample	Doses for 98 sample
KASP assay mix	Allele-specific forward oligo-primer	1 µl	98 µl
	Allele-specific forward oligo-primer	1 µl	98 µl
	Common reverse primer	1 µl	98 µl
	KASP master reaction mix	5 µl	490 µl
	MQ	1 µl	98 µl
Total	Master mix	9 µl	882 µl
	DNA template	1 µl	1 µl x 98
Total reaction volume		10 µl	10 µl x 98

For non-template control, 1 µl of MQ was added instead of DNA template

3 RESULTS

3.1 Analysis of original NGS read and sample quality

The NGS whole genome sequencing produced 200M reads in PR, BR, and BS, respectively (Table 2). Most of the quality parameters for reads of both the bulks and PR generated by MultiQC were found satisfactory. However, half of the samples in both bulks and PR showed a warning in sequence quality histogram signifying that most of the Phred Scores are in the green zone (Phred Score > 30 except for last few base pairs). The GC content in all samples was failed in BR (Appendix IV). To have quality reads, the reads were trimmed for each bulk and PR. The trimmed reads were used to derive list of k-mers. The k-mers which were represented once, accounting ~5% of total volume of k-mers were removed because they probably represent sequence errors. The remaining error-free k-mers (freq>1) were used to construct k-mers frequency histogram for each bulk and PR. The k-mers frequency histogram showed 2 peaks in PR and BS, whereas four peaks were found in BR (Figure 5). The second peak contained most of the k-mers: more than 85% unique k-mers and more than two-thirds total k-mers (Appendix V).

Table 2 Read and k-mers information for resistant parent, resistant bulk and susceptible bulk

Parent or bulks	Yield (Gb)	Reads count	K-mers including read errors		Error-free k-mers (k>1)		Median frequency
			Unique k-mers	Total k-mers	Unique k-mers	Total k-mers	
PR	18.98	196,457,428	1,693,996,440	20,595,198,060	724,223,720 (57.25%)	19,625,425,340	24
BR	23.18	237,565,974	225,063,5764	24,748,151,032	1,029,485,093 (54.26%)	23,527,000,361	21
BS	17.89	182,933,860	1,965,215,375	18,607,551,517	982,411,847 (50.01%)	17,624,747,989	17

BR and BS contain two peaks for which median coverage was determined by taking average between them; PR = resistant parent; BS = susceptible bulk; BR = resistant bulk

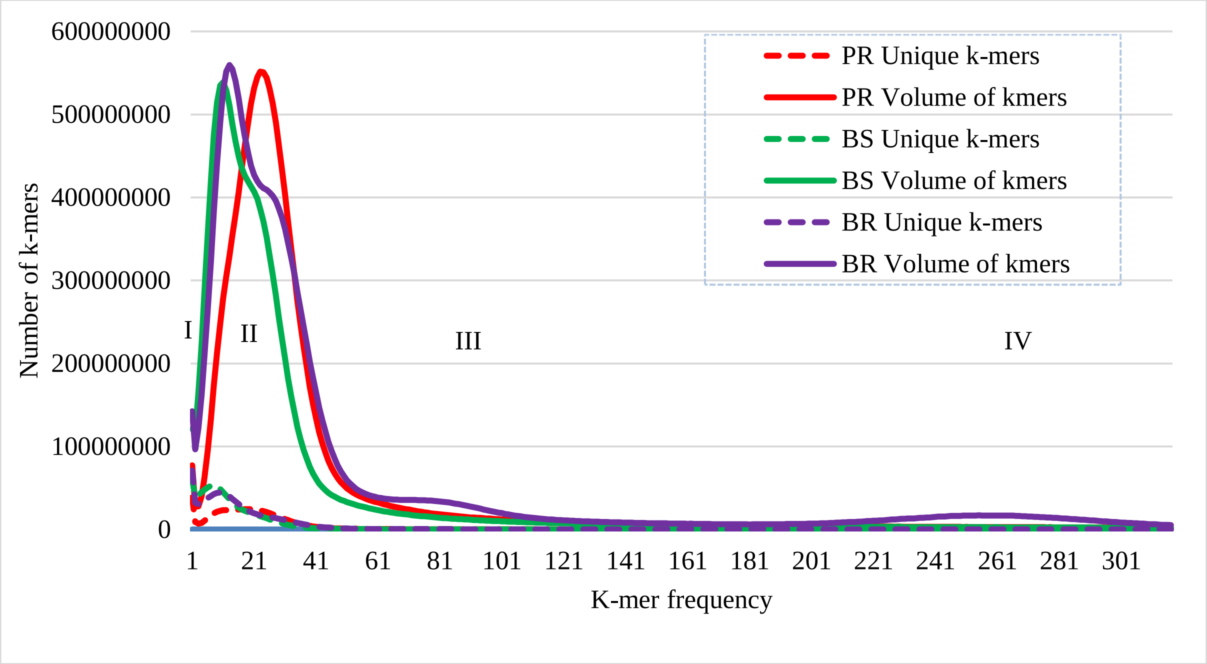


Figure 5 The k-mers histogram for resistant parent, resistant bulk and susceptible bulk; PR = resistant parent; BR = resistant bulk; BS = susceptible bulk; the roman letter I, II, III, and IV represent the observed histogram peak including k-mers frequency of 2 to 3, 4 to 50, 51 to 100 and 200-320 respectively

3.2 Bulked segregant k-mers analysis (BSKA)

3.2.1 Resistance bulk-specific k-mers selection using different BSKA approaches

The 1:1 segregation ratio implied that the single dominant resistance gene was responsible for target trait. The PR has two haplotypes *i.e.* resistance haplotype represented as R and susceptible haplotype represented as r in the target region of the genome. Susceptible parent contains two other S haplotypes, like haplotype r' and haplotype r'' in the same region of a genome. Accordingly, the BR should contain R, r' and r'' haplotypes whereas BS should contain r, r' and r'' haplotypes (Appendix VI). We used above-mentioned haplotype concept to deduct haplotype-specific k-mers list. First, we deduced a BR specific k-mers list which was present in BR or PR or both but not present in BS using five different BSKA approaches as described in Table 3.

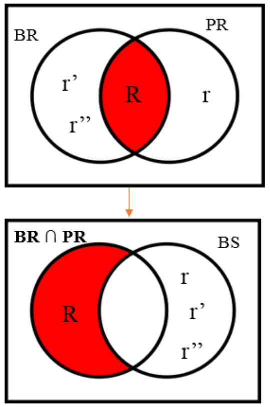
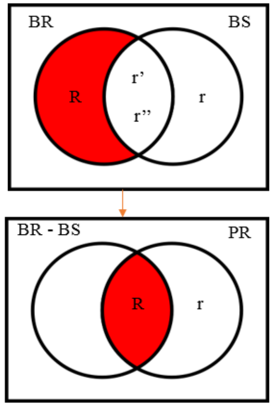
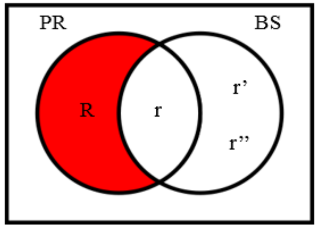
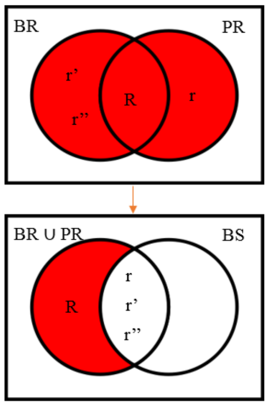
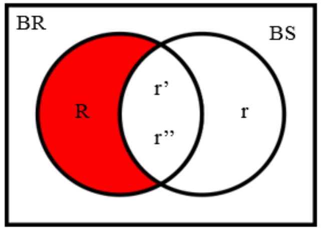
In BSKA approach I, we first produced a list of common k-mers present in PR and BR (BR ∩ PR) (unique k-mers = 674,768,948, total k-mers = 15,297,710,202). The BS was subtracted from BR ∩ PR to keep common k-mers that are present in PR and BR but not in BS ((BR ∩ PR) – BS). Finally, we obtained 18,752,497 unique and 102,032,311 total k-mers. BSA approach II is also similar to BSA approach I. The k-mers present in BR but not present in BS (BR-BS) was derived which was followed by listing shared k-mers among BR-BS and PR. Both the methods resulted in a same number of k-mers count specific to BR (Table 3).

BSA approach III used only PR and BS-specific k-mers list during analysis. This approach is quite straight-forward: the k-mers present in BS was directly subtracted from PR *i.e.* PR-BS. We obtained 8,717,733 unique k-mers and 123,256,247 total k-mers associated with BR. To hold most of the k-mers in the final list, we developed BSA approach IV which was like approach I except the use of union function between PR and BR (BR ∪ PR) instead of using intersection during the first step of the analysis. We found 1,078,939,865 unique k-mers and 43,152,425,701 total k-mers that were present in both BR and PR. The BS was then subtracted form BR ∪ PR. As a result, BR retained 132,576,297 unique k-mers and 2189925627 total k-mers, which were the highest number of k-mers among all derieved approaches. Moreover, we developed BSA approach V which is based on analysis of k-mers associated with BR and BS. We directly produced BR specific k-mers list that were present in BR but not present in BS by subtracting BS from BR *i.e.* BR-BS. We obtained 27,951,629 unique and 1,769,582,641 total BR specific k-mers list (Table 3).

3.2.2 Bulk specific k-mers mapping to reference genome DM

To avoid the most likely error-prone k-mers (false positives), we produced the list of k-mers with coverage threshold (cut-off) 6, 8, 10, 12, 14 and 16 for each BSKA approach. Each k-mers list with different coverage threshold was then mapped to the reference genome DM (Potato Genome Sequencing Consortium, 2011) and k-mers count per 1Mb region on all potato chromosomes was recorded. Surprisingly, we obtained comparable numbers of k-mers mapped to the chromosome 11 and chromosome 12. In BSKA approach I and BSKA approach II, we found the higher peak at chromosome 12 followed by chromosome 11 at coverage threshold 6 and 8 whereas at coverage threshold 10, 12, 14 and 16 the chromosome 11 has the higher peak than chromosome 12 (Appendix VII). In BSKA approach III, we found a similar trend as with BSA approach I, however, chromosome 12 recorded the highest peak of k-mers count than chromosome 11 at all designed coverage threshold (Appendix VIII). In BSKA approach IV, the highest k-mers peak was observed on chromosome 12, however, chromosome 11 also recorded comparable peak (Appendix IX). In BSKA approach V, the count of k-mers after mapping to reference genome DM showed the highest number of mapped k-mers at chromosome 11 followed by chromosome 12 except cut-off 6 (Appendix X).

Table 3 Different BSKA approaches used to derive resistance bulk-specific k-mers list from intraspecific crosses of *S. capsicibaccatum* using NGS data

	Approach I: BR specific k-mers which are shared with PR but not present in BS	**Approach II:** BR specific k-mers not present in BS but coming from PR	**Approach III**: PR specific k-mers but not present in BS (most likely goes to BR)	**Approach IV:** BR and PR specific k-mers but not present in BS	**Approach *V*:** BR specific k-mers but not present in BS
Input NGS data	BR, BS, PR	BR, BS, PR	BS, PR	BR, BS, PR	BR, BS
	 BR PR r' R r r'' BR ∩ PR BS R r r' r''			 BR PR r' R r r'' BR ∪ PR BS R r r' r''	
Mathematical description	BR ∩ PR (BR ∩ PR) – BS	BR – BS (BR – BS) ∩ PR	PR – BS	BR ∪ PR (BR ∪ PR) – BS	BR – BS
K-mers count (error free)	Unique = 18,752,497 Total = 102,032,311	Unique = 18,752,497 Total = 102,032,311	Unique = 8,717,733 Total = 123,256,247	Unique = 132,576,297 Total = 2,189,925,627	Unique = 27,951,629 Total = 1,769,582,641
Remarks	Less unique k-mers, however, select against contaminants in BR	Less unique k-mers, however, select against contaminants in BR	Least unique k-mers, this approach does not include BR, therefore, might not contain contaminants from BR	The highest number of k-mers list, however, select contaminants too	Second highest number of k-mers list

Two haplotypes, resistance (R) and susceptible (r) from resistant parent and two susceptible haplotypes (r' and r'') from the susceptible parent. The resistant bulk should contain R, r' and r'' haplotypes whereas susceptible bulk should contain r, r' and r'' haplotypes; BSKA = bulked segregant k-mers analysis; PR = resistant parent; BR = resistant bulk; BS = susceptible bulk; NGS = next-generation sequencing, R = resistant; S = susceptible; the red area represents output. The Venn diagram was taken from https://en.wikipedia.org/wiki/Venn_diagram and modified according to inheritance based haplotype deduction concept in the present study. Intraspecific crossing represents the crossing of *S. capsibaccatum* resistance accession and *S. capsibaccatum* susceptible accession.

For chromosome 11, the highest peak was observed on 0-2 Mb region (Figure 6 as a reference example derived from BSKA approach I), which is expected position of the *Rpi-cap1* gene. In chromosome 12, the highest peak was observed in 58-61 Mb chromosome region. The peak in chromosome 12 was artefacts. Most likely, the small bulk sizes (11 plants for each bulk) used in whole genome sequencing have accidentally identified two genomic regions. Besides, we also got artefact peaks in chromosome 7 and chromosome 3 (Appendix VII; Appendix VIII; Appendix IX; Appendix X), however, these peaks were not sufficiently high (under-represented) inferring fewer than 11 plants shared these regions.

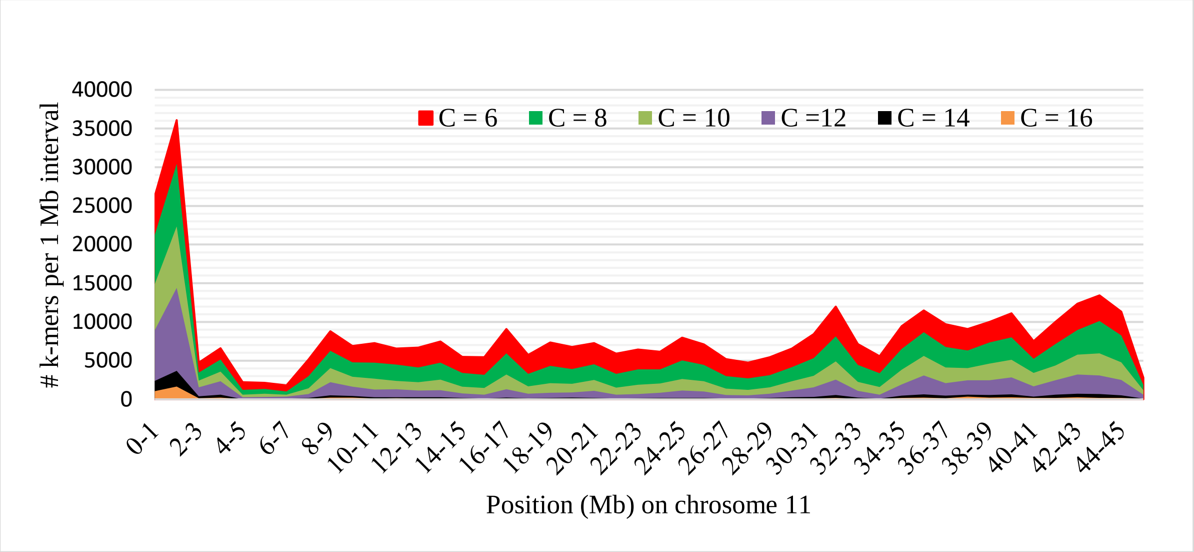


Figure 6 Resistant bulk specific k-mers which are shared with resistant parent but not present in susceptible bulk (BSKA approach I) on chromosome 11 at different coverage threshold; C represents the coverage threshold, for example, C = 6 holds all the k-mers with coverage depth 6 and above

3.2.2.1 Determining region of interest on unanchored chromosome

Reference genome of potato covered 623 Mb sequences anchored to chromosomes and 17 Mb of unanchored scaffolds (Potato Genome Sequencing Consortium, 2011). While mapping wart disease R haplotype-specific k-mers to reference genome DM, Charlotte Pradhomme (unpublished data) found a k-mers peak on chromosome 0, on 20750 Kb to 20950 Kb bin interval. She showed that this unanchored scaffold belonged to chromosome 11, 0-4 Mb region. Both these Chromosome11 and Chromosome 0 regions harbour TNL genes. We hypothesized this region is also associated with *Rpi-cap1* gene or at least *NB-LRR* region.

In the present study, there was a peak on 20-21 Mb region of chromosome 0 linked with a peak on chromosome 11 target region (0-2 Mb region). We zoomed out in to that region and counted the number of k-mers mapped per 50 Kb interval of potato chromosome 0. Interestingly, there was the highest number of mapped k-mers on 20800 kb to 20950 kb interval of potato chromosome 0 (Figure 7) in all above described BSA approaches except BSA approach III. During haplotype-specific k-mers selection, we chose k-mers mapped to 20800 Kb to 20950 Kb region of chromosome 0 with k-mers mapped to 0-2 Mb region of chromosome 11 for downstream analysis.

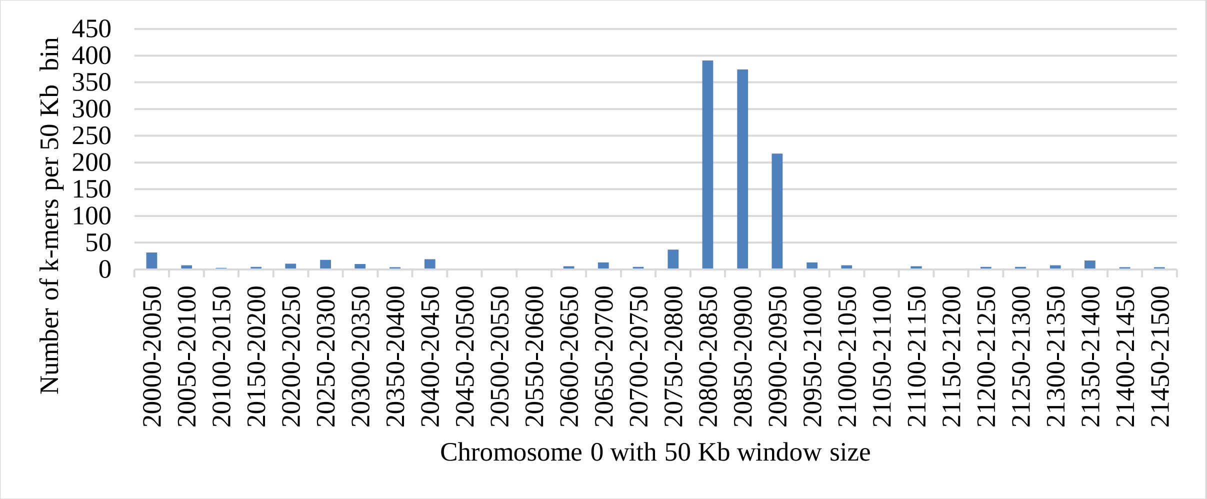


Figure 7 Number of mapped k-mers per 50 Kb interval on chromosome 0; this graph is exemplary graph using BSA approach IV with coverage threshold 16 and above

3.2.3 Selection of the BSKA approach for optimal identification of QTL peaks

3.2.3.1 Selection of the BSKA approach with the highest signal to noise ratio and the role of coverage threshold

Verzaux et al. (2012) reported markers that flank the *Rpi-cap1* gene and are in a long arm of chromosome 11 (Figure 2), the hot spot of *R* gene where a significant volume of unique k-mers produced in the present study was mapped to expected region (0-2 Mb region) with a good signal-to-noise ratio (SNR). We found increased SNR on target region on increasing coverage threshold in all BSKA approaches except BSKA approach III, to which the most important constituent of the present study – BR - was not used during the analysis. On increasing threshold, the k-mers from single copy genes most likely retained in the final list, while k-mers from multicopy would be removed. In general, the BSKA approach V recorded the highest SNR, which was followed by BSKA approach I or II, BSKA approach IV and least SNR was observed in BSKA approach III (Figure 8).

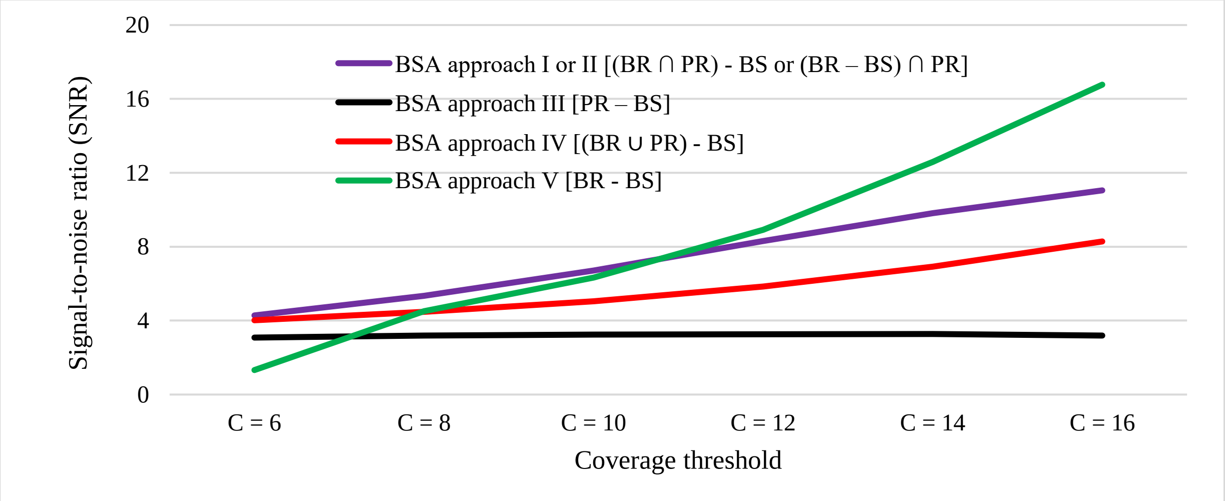
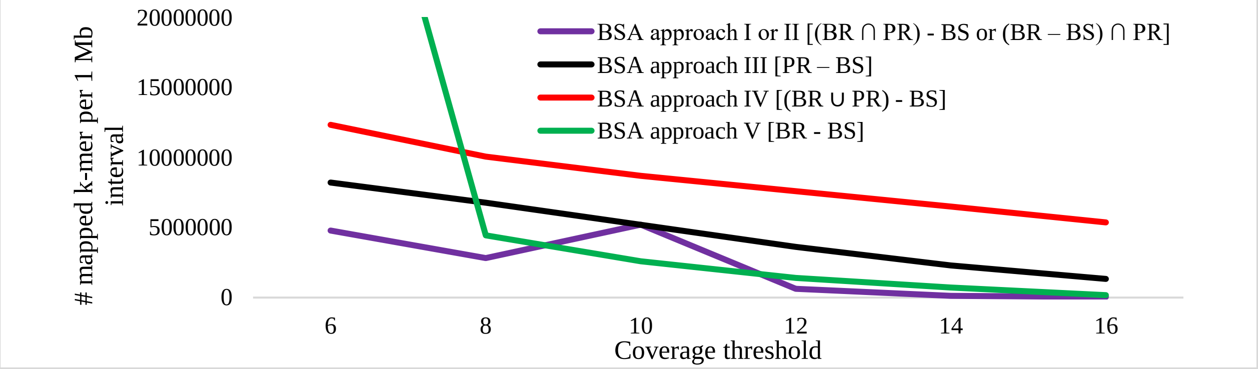
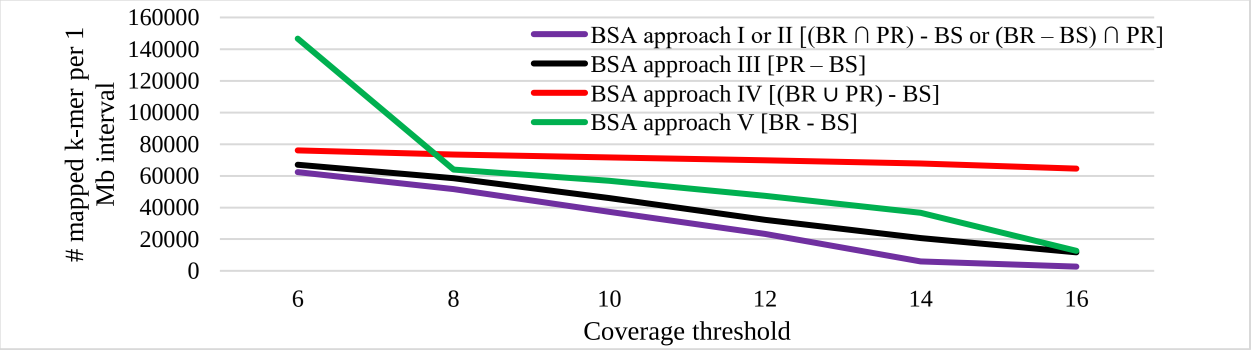


Figure 8 Signal-to-noise ratio (SNR) observed on chromosome 11, 0-2 Mb region using different BSKA approaches; C = coverage threshold

We compared the count of mapped k-mers per 1 Mb interval on all potato chromosome and on the target region of chromosome 11 as well. Except for coverage threshold 6, union-based BSA approach IV which list all the k-mers that are present in both PR and BR but not present in BS retained the highest number of mapped k-mers on all potato chromosome, which was followed by BSA approach III. BSA approach I or II and BSA approach V retained least number of k-mers mapped to all potato chromosome (Figure 9). Similarly, we found BSA approach IV retained the highest number of k-mers mapped to target region of chromosome 11, which was followed by BSA approach V, BSA approach III and BSA approach I, respectively. Considering SNR and number of mapped k-mers on all potato chromosome and the target region of chromosome 11, BSA approach IV was chosen for downstream analysis.



a. Number of mapped k-mers per 1 Mb bin interval for all potato chromosomes

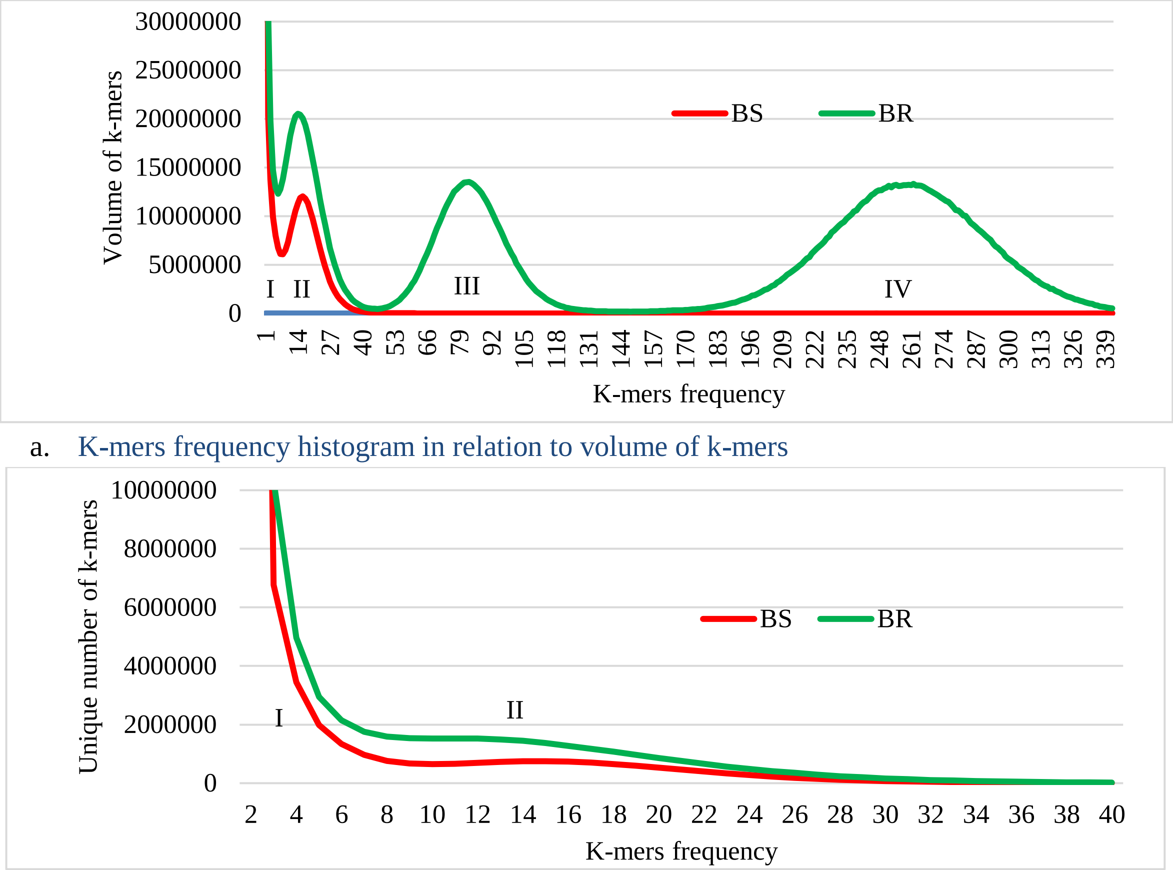


b. Number of mapped k-mers per 1 Mb bin interval for chromosome 11 target region

Figure 9 Number of mapped k-mers per 1 Mb interval in all potato chromosome and chromosome 11, 0-2 Mb region; PR = resistant parent; BR = resistant bulk; BS = susceptible bulk

3.2.3.2 Selection of k-mers representing single copy gene from BSKA approach IV

The candidate gene region (peak) observed in present study is very broad which most likely diluted the signal due to broad bin size. Therefore, to derive k-mers list representing single copy gene, we selected BR specific histogram from BSA approach IV despite its intermediate performance on achieving SNR. The k-mers histogram of BR showed four peaks in k-mer frequency (Figure 10a). The k-mers that occurred twice represented 7% of total k-mers, however, they accounted for 60% of the unique k-mers. The first peak (2-6 frequency window) contains ~74% unique k-mers (Figure 10b), however must of them are an error. The second peak (7-32 frequency window) contains ~18% of unique k-mers, however, they were BR-specific. Here, we might expect median coverage as the half of sum of PR and BR (Figure 5), and found accordingly. The third (52-122 frequency window) and fourth peak (182-342 frequency window) recorded very low percentage of unique k-mers among total unique k-mers (less than 5% on each) (Figure 10b).



a. K-mers frequency histogram in relation to volume of k-mers

b. K-mers frequency histogram in relation to unique k-mers

Figure 10 Resistance and susceptible bulk specific k-mers histogram derived using BSKA approach IV; volume of k-mers = amount of unique k-mers with specific coverage x occurrence of respective k-mers; k-mers frequency = k-mer coverage or depth of respective k-mer, the roman letter in the figure represents the respective peak; BR = (k-mers list specific to) resistance bulk; BS = (k-mers list specific to) susceptible bulk

The k-mers from third and fourth peak were mapped to reference genome DM. No k-mers from fourth peak mapped to the reference genome, and only 950 kmer from third peak (0.02% among 5,216,488 k-mer) mapped to reference genome. This result suggested that peak III and IV were caused by a contaminants coming from BR (Refer Figure 5). The blasting of these unmapped k-mers into NCBI database showed they are part of the bacterial and fungal genome. Among the 950 k-mers from peak 3, 125 unique k-mers mapped to 1-2 Mb bin of reference genome in chromosome 11. This suggested that these were derived from high copy number potato sequences, and not from the contaminating microbes.

To retrieve unique k-mers without multicopy sequences, lower and upper coverage thresholds such as 7 to 23, 7 to 30, 7 to 35, and 7 to 40 were set and evaluated for further downstream analysis. Each k-mers list with these coverage thresholds settings was then mapped to reference genome DM followed by k-mers counting per 1 Mb region of potato chromosome 11. We did not compare for chromosome 0. We found there was increasing SNR in our target region of chromosome 11 on increasing upper coverage indicating that the coverage above 23 contributes more on a signal on our target region than the untargeted region (Table 4). Increasing signal is also associated with higher probability of selecting multicopy sequences or paralogs of a target gene. Therefore, to avoid paralogs and to select unique or single copy marker (we prefer quality over quantity), we chose k-mers with coverage depth 7 to 23 as optimum for downstream analysis. It was also supported by median coverage of BR specific peak which was around 17 (second peak in Figure 10a).

Table 4 Signal-to-noise ratio observed in chromosome 11, 0-2 Mb region using resistance bulk specifik k-mers list of different coverage threshold

Coverage depth	Signal mean	Noise mean	SNR
6 and above	38,033	15836	2.40
7 and above	37,341	14,270	2.62
7 to 23	16,991	11,542	1.47
7 to 30	28,916	13,451	2.15
7 to 35	34,247	14,005	2.45
7 to 40	36,298	14,190	2.56

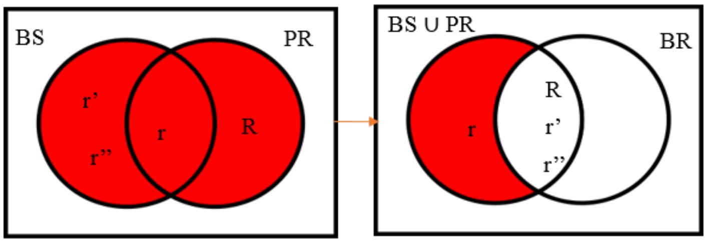
SNR = Signal-to-noise ratio; signal mean represents k-mers average count in 0-2 Mb region whereas noise mean represents average count in rest of the region of chromosome 11 when all k-mers mapped to the potato reference genome

3.2.3.3 Susceptible bulk specific k-mers selection and mapping to reference genome DM

To identify reads and variant from the S haplotype, first, we derived all error-free k-mers (N unique = 85,503,051, N total = 375,224,063) present in BS and PR but not present in BR ((BS ∪ PR) – BR) (Table 5) and the histogram was drawn. Unlike BR specific k-mers list ((PR U BR) –BS), BS specific k-mers list only kept first and the second peak when k-mers were mapped to the referencce genome followed by k-mers counting per1 Mb bin of potato chromosome. It lacks the third and fourth peak (Figure 10a). The BS specific k-mers list retained 17.13% less k-mers volume than BR specific k-mers list. Also, it retained 35.51% lower unique k-mers than BR specific k-mers list. This reduction in unique and total k-mers could be explained by an absence of the third and fourth peak in the BS specific k-mers list. More than 85% unique k-mers were present in the first peak, however, the second peak only kept 12% unique k-mers (Figure 10b). Still, the second peak kept 46% total k-mers in terms of k-mers volume or total k-mers.

The second peak became our target peak. It also resembled the second peak in BR specific k-mers list with almost same median value (17). To retrieve unique k-mers without multicopy sequences, the same lower and upper coverage threshold was set (7 to 23) as optimum for downstream analysis.

Table 5 BSA approach used to derive susceptible bulk-specific k-mers from intraspecific crosses of diploid *S. capsicibaccatum* using NGS data on resistant bulk, susceptible bulk and resistant parent

	BS specific k-mers which are shared with PR but not present in BR
Input NGS data	PR, BR, BS
	
Mathematical description	(BS ∪ PR) (BS ∪ PR) – BR
K-mers count (error free)	N unique = 85,503,051 N total = 375,224,063

Intraspecific crossing represents the crossing of *S. capsibaccatum* resistance accession and *S. capsibaccatum* susceptible accession

3.2.4 Estimation of SNP frequency

The BR specific k-mers list with coverage depth 7-23 retained 21M unique k-mers. Only 44.48% unique k-mers mapped to the entire reference genome DM. There were 36,404 k-mers mapped to the target region (chromosome 11, 0-2 Mb region and chromosome 0, 20800 Kb to 20950 Kb region) and therefore considered as R haplotype-specific k-mers list. Similarly, BS specific k-mers list with threshold 7 to 23 retained 11M unique k-mers. Among them, 61.83% of k-mers mapped to the entire reference genome DM. We found 50,539 k-mers mapped to the target region and therefore considered as S haplotype-specific k-mers list (Table 6).

Table 6 K-mers details specific to resistance and susceptible haplotype with coverage threshold 7 to 23

	Count	
	Resistance bulk/haplotype	Susceptible bulk/haplotype
K-mers retained	21,114,437 (unique k-mers) 288,243,438 (total k-mers)	11,076,109 (unique k-mers) 156,177,384 (total k-mers)
Unique k-mers mapped to entire potato chromosomes	9,392,373	6,848,318
Unique k-mers mapped to chromosome 11, 0-2 Mb region	33,982	46,657
Unique k-mers mapped to chromosome 0, 20800-20950 Kb region	2424	3882
Number of putative SNPs present on region of interest on chromosome 11 and 0	1,174 (1 SNP per 1,831 bp interval)	1,630 (1 SNP per 1,319 bp interval)

Based on a count of k-mers that mapped to a target region of the genome, we estimated the putative number of SNPs. We hypothesized one haplotype-specific SNP produced maximally 31 mapped k-mers (two or more SNPs within 31 bp would reduce the kmer to SNP ratio and hence SNP frequency). Therefore, the maximum number of SNPs in a particular region of the chromosome would be the total number of unique k-mers mapped to the selected region divided by the length of k-mers. According to these calculations, we found 1096 and 1505 SNPs on a target region of chromosome 11 for R haplotype and S haplotype, respectively. Again, we found 78 and 125 SNPs one target region of chromosome 0 for R haplotype and S haplotype, respectively (Table 6). Still, the coverage threshold largely affects the number of retrieved k-mers and it would then affect the expected number of SNPs in the target region. The number of SNPs is very low to cover the target region.

3.3 Haplotype specific read retrieval and *de-novo* assembly

The lower number of SNPs implied that the contiguous assembly in the target region is not possible. Still, identifying SNPs and developing primers using k-mers mapping to the reference genome might miss important SNPs list, if they are not really matched to the reference genome. To determine the SNPs that are specific to resistance gene but not to the multicopy, haplotype-specific *de-novo* assembly was planned for downstream analysis. The selected k-mers from above-explained procedures were first filtered using mapping quality criteria. Among 36,406 unique k-mers associated with R haplotype, 77.28% unique k-mers passed the mapping quality criteria (q = 2) (Supplementary file 2, Folder name: R haplotype specific k-mers list 7-23, File name: R haplotype 7 to 23.kmer). We retrieved the reads that contain at least 1 kmer from both PR and BR. The R haplotype retained 35,915 total R1 and R2 reads (Table 7). Moreover, the same procedure was followed to retrieve reads associated with S haplotype. Among 50,499 k-mers associated with S haplotype, 79.94% reads passed the quality criteria (q = 2) (Supplementary file 3, Folder name: S haplotype specific k-mers list 7-23, File name: S haplotype 7 to 23.kmer). We retrieved the reads that contain at least 1 kmer from both PR and BS. We got each 42181 R1 and R2 reads associated with S haplotype (Table 7). Again, the R1 and R2 reads associated with each haplotype were trimmed and the unpaired and other reads that do not meet the basic quality criteria were removed. The significant numbers of reads specific to R haplotype were trimmed in comparison to S haplotype (Table 8). Both trimmed reads and untrimmed reads (Supplementary file 4, Folder name: Haplotype specific reads, File name: R haplotype specific reads, S haplotype specific reads) were used for haplotype-specific *de-novo* assembly and comparison was made in order to select the best assembly for downstream analysis.

Table 7 K-mers and reads information for resistance and susceptible haplotype

Reads or k-mers	Count	
	Resistance haplotype	Susceptible haplotype
Unique k-mers	36,406	50,499
Quality passed k-mers	28136	40,367
Retrieved reads from PR	18,208 each R1 and R2	25,663 each R1 and R2
Retrieved reads from BR or BS	17,707 each R1 and R2	16,518 each R1 and R2
Total read	35,915 each R1 and R2	42,181 each R1 and R2

PR = resistant parent; BR = resistant bulk, BS = susceptible bulk; R1 = forward reads; R2 = reverse reads; both haplotypes were 7 to 23 coverage depth

Table 8 Reads associated with resistance and susceptible haplotype before and after trimming

Reads type	Resistance haplotype		Susceptible haplotype	
	Forward reads (R1)	Reverse reads (R2)	Forward reads (R1)	Reverse reads (R2)
Total read (Input)	35,915	35,915	42,181	42,181
Untrimmed read (quality passed read)	25,883	25,883	39,051	39,051
Trimmed unpaired read	1,892	122	2,849	194

The insert size for R haplotype reads was calculated to be 342.52 with standard deviation 51.32. Similarly, the insert size for trimmed reads was determined which was almost similar. The both trimmed and untrimmed reads were assembled using SOAPdenovo2 followed by assembly evaluation. We found better assembly result in the latter case; therefore, untrimmed reads were used for final assembly. The decrease in assembly characteristics during the use of trimmed reads is most likely due to the missing of some important reads that, otherwise would efficiently help to fill the gaps, for example, unpaired reads. The use k-mer sizes of 41, 51, 61, 71, and 81 for *de-novo* assembly were compared. The assembly evaluation showed that the k-mer size 51 produced the highest size of the scaffold, median scaffold size, longest scaffold and longest contig. The assembly result using kmer size 61 produced better N50 scaffold length and scaffold count. Again, the k-mer 81 produced the best N50 contig length and L50 contig count (Table 9). It is apparent that using different k-mers sizes produced results with varying assembly characteristics. To trade-off the balance, we have chosen *de-novo* assembly using a k-mer size of 51 for contig based mapping and contig based SNPs mining. We also assembled reads specific to S haplotype and found k-mer size 51 was optimum to produce better S contigs. The result is not presented here as *de-novo* assembled S haplotype read was not used in downstream analysis.

Table 9 Assembly characteristics evaluation for *de-novo* assembled resistance haplotype produced using different k-mer sizes

Assembly characteristics	K-mer sizes				
	41	51	61	71	81
Number of scaffolds	5,623	1,394	1,260	1,149	934
Total size of scaffold	1,529,391	917,643	852,132	763,927	631,950
Longest scaffold	2,748	8,210	7,670	7,670	6,640
Shortest scaffold	100	100	100	112	132
Number of scaffolds > 500 nt	871	781	696	580	437
Number of scaffolds > 1K nt	145	195	177	162	147
Mean scaffold size	272	658	676	665	677
Median scaffold size	151	563	547	504	487
N50 scaffold length	427	758	763	749	802
L50 scaffold count	982	338	308	269	208
N50 scaffold - NG50 scaffold length difference	427	758	763	749	802
Percentage of assembly in scaffolded contigs (%)	18.70	31.40	32.40	15.10	15.40
Percentage of assembly in unscaffolded contigs (%)	81.30	68.60	67.60	84.90	84.60
Number of contigs	6,077	1,790	1,618	1,290	1,029
Number of contigs in scaffolds	849	748	678	269	184
Number of contigs not in scaffolds	5,228	1,042	940	1,021	845
Total size of contigs	1,505,399	897,949	835,100	757,139	627,746
Longest contig	2,707	8,035	7,670	7,670	6,640
Shortest contig	13	97	100	112	132
Number of contigs > 500 nt	664	601	559	534	414
Number of contigs > 1K nt	108	174	152	154	141
Mean contig size	248	502	516	587	610
Median contig size	151	351	372	452	450
N50 contig length	287	707	701	717	740
L50 contig count	1,224	368	348	289	227
N50 contig - NG50 contig length difference	287	707	701	717	740

Refer Supplementary file 5, file name: R haplotype k-51 denovo result.scafseq for contigs and scaffolds sequences

3.4 Read Mapping to reference genome and haplotype-based variant detection

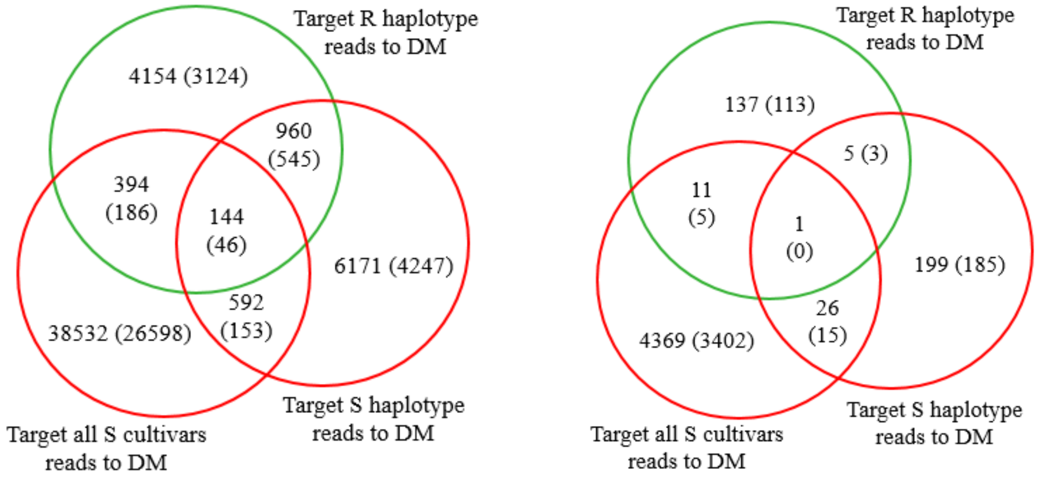
To locate the *Rpi-cap1* locus to reference genome DM, the R and S haplotype-specific trimmed R1 and R2 reads were mapped to reference genome DM using BWA-mem read aligner. Among 51,766 R haplotype-specific trimmed reads retrieved above, 78.11% reads were paired properly. However, we found higher number reads mapped to the reference genome DM implying reasonable reads whose mate not paired properly or mate mapped to a different chromosome (Table 10). We found 39,452 reads mapped to the region of interest in chromosome 0 and chromosome 11 (Appendix XI). Similarly, among 78,102 R1 and R2 reads paired in sequencing in S haplotype, 82.06% reads were properly paired. Again, quite a large number of reads were mapped to the reference genome implying a significant number of reads whose mate mapped to a different chromosome or mate pairs not mapped properly (Table 10). There were 66,490 reads mapped to the region of interest (Appendix XII). The higher mapping reads in both cases may imply that the reads of transposons or paralogs are present in the analysis product or they may be mapped to other *NB-LRR* sites. Still, the significant number of singletons retained in both R and S specific haplotype read mapping implying sequencing error or something else. Moreover, the trimmed reads R1 and R2 of each nine susceptible cultivars mapped to reference genome DM. The mapping produced alignment result in BAM format.

Table 10 Read and mapping characteristics of resistance and susceptible haplotype-specific reads to reference genome DM

Read and mapping characteristics	Reads count	
	Resistance haplotype-specific reads mapping to DM	Susceptible haplotype-specific reads mapping to DM
Reads paired in sequencing (input)	51,766 (25,883 R1 and 25,883 R2)	78,102 (39,051 R1 and 39,051 R2)
QC-passed reads	55,170	82,290
Mapped reads to DM	54,678 (99.11%)	81,831 (99.44%)
Properly paired reads	40,436 (78.11%)	64,094 (82.06%)
Reads with itself and mate mapped	50,860	77,226
Singletons reads	414 (0.80%)	417 (0.53%)
Reads with mate mapped to a different chromosome (MapQ>=1)	7,604	9,514
Reads with mate mapped to a different chromosome (mapQ>=5)	4,269	5,655

The higher number of QC-passed reads than provided input reads (reads paired in sequencing) represents a mapping of those reads in more than 1 places. It was reflected as mapping quality during visualization in IGV

Following read mapping, variants associated with R haplotype, S haplotype, and nine susceptible cultivars were detected using FreeBayes relative to the reference genome DM for the target region of both chromosome 11 and chromosome 0. The higher number of variant (6,370) was found in S haplotype than R haplotype (4,291) relative to reference DM (Supplementary file 6, Folder name: DM based haplotype-specific variants, File name: Variants on chromosome 11 and chromosome 0). FreeBayes variant list contained SNPs, InDels, multi-nucleotide polymorphisms (MNPs) and composite insertion and substitution events. We filtered and kept only SNPs in the final list. We found 75% of the total R haplotype-specific variants were SNPs (Figure 11; Supplementary file 7, Folder name: DM based haplotype-specific SNPs, File name: SNPs on chromosome 11 and Chromosome 0).


Target R haplotype reads to DM
4154 (3124)
960 (545)
394 (186)
144 (46)
6171 (4247)
592 (153)
38532 (26598)
Target S haplotype reads to DM
Target all S cultivars reads to DM
Target R haplotype reads to DM
137 (113)
5 (3)
11 (5)
1 (0)
199 (185)
26 (15)
4369 (3402)
Target S haplotype reads to DM
Target all S cultivars reads to DM

a. Variant observed on chromosome 11

b. Variant observed on chromosome 0

Figure 11 Variants in the target region of chromosome 11 and chromosome 0 for each resistance specific haplotype, susceptible specific haplotype and all susceptible cultivars relative to DM; the value in parenthesis indicates the number of SNPs; S cultivars represents the susceptible cultivars

3.5 Read Mapping to *de-novo* assembled contigs and haplotype-based variant detection

The trimmed reads associated with each R haplotype, S haplotype and other nine susceptible cultivars were mapped to *de-novo* assembled reference R contigs (k-mers size =51). For R haplotype, among 51,766 reads that paired in sequencing, 97.38% reads were mapped to R contigs. 83.76% reads were paired properly. There was 2.58% singleton and 10.70% reads with mate mapped to different contigs implying scaffolding has not been performed in a maximal way during *de-novo* assembly. Similarly, among 78,102 reads paired in sequencing associated with S haplotype, 45.57% reads paired properly. There were 66.14% reads mapped to R contigs. There were 11.50% reads whose mate mapped to different contig and the singleton count was rather higher (7.86%). There was less mapping percent when S haplotype-specific reads mapped to R contig, however, the final count that mapped to R contig turn out to be same for both haplotypes (Table 11). Other nine susceptible potato cultivars were also mapped to R contigs.

Table 11 Read and mapping characteristics of resistance and susceptible haplotype-specific reads to *de-novo* assembled resistance contigs

Read and mapping characteristics	Reads count	
	Resistance haplotype-specific reads mapping to R contigs	Susceptible haplotype-specific reads mapping to R contigs
Reads paired in sequencing (input)	51,766 (25,883 R1 and 25,883 R2)	78,102 (39,051 R1 and 39,051 R2)
QC-passed read	52,924	80,192
Mapped reads to R contigs	51,536 (97.38%)	53,035 (66.14%)
Properly paired reads	43,360 (83.76%)	35,592 (45.57%)
Reads with itself and mate mapped	49,040	44,810
singletons read	1,338 (2.58%)	6,135 (7.86%)
Reads with mate mapped to a different chromosome	5,534	8,992
Reads with mate mapped to a different chromosome (mapQ>=5)	5,241	8,046

Following alignment, the variant calling was performed for R haplotype, S haplotype and other nine susceptible cultivars (S cultivars) in relative to R contig. There were 15 variants on R haplotype reads in relative to R contig inferring error variant, however, they may be from the repetitive regions. There were 3,722 variants associated with S haplotype reads in relative to R contigs (Supplementary file 8, Folder name: R contigs based haplotype-specific variants, File name: Variants from R and S haplotype). Again, multiple BAM file of 9 cultivars showed 34,139 variants. Among the list of variants, 62.20% and 74.20 % variant were SNPs for both S haplotype and reads from nine susceptible cultivars in relative to reference R contigs, respectively (Table 12; Supplementary file 9, Folder name: R contigs based haplotype-specific SNPs, File name: SNPs from R haplotype and S haplotype).

Table 12 Variant for resistance specific haplotype, susceptible specific haplotype and all susceptible cultivars relative to reference *de-novo* assembled R contigs

	Resistance haplotype-specific reads	Susceptible haplotype-specific reads	Reads from nine susceptible cultivars
Variant	15	3,722	34,139
SNPs	10	2,315	25,332

In order to determine the position of R contig to reference genome DM, we anchored the assembled R haplotype-specific contigs and scaffolds to the reference genome DM using BLAST searches with certain criteria (eg. maximum target sequence = 1, e-value = 1e-16) that gave the most likely position of contigs in reference DM. We selected only those contigs that anchored to 0.8 Mb region to 1.25 Mb region of chromosome 11 and 20800 Kb to 20950 Kb region of chromosome 0. There were 64 contigs anchored to chromosome 0 and 790 contigs anchored to chromosome 11. Among them 37 contigs anchored to the target region (20800 Kb to 20950 Kb) of chromosome 0 and 288 contigs anchored to the target region (0 to 2 Mb) to chromosome 11. Among 325 contigs (37+288), there were 166 (15 + 151) contigs without any variant in associated mapped reads from S haplotype to R contig, therefore rejected. Finally, there were only 22 (37-15) contigs that anchored to chromosome 0 and 237 (288-51) contigs that anchored to chromosome 11 region of interest and was associated with variant (Appendix XIII).

3.6 Primer design for KASP assay

The contigs that anchored to the region of interest in reference genome DM and retained SNPs relative to the *de-novo* assembled reference genome (R contig) were visualized in IGV using mapped BAM files. The SNPs was only selected if both resistance and susceptible specific reads flanking the SNPs had optimum read coverage and that specific SNPs was not present in other 9 susceptible variety reads. The primers were designed for KASP genotyping using specific *de-novo* assembled anchored contigs. Considering time and resources we had, design and validation of KASP markers were prioritized and done. Based on the position of the previously described two flanking markers such as Cp58 (0.814 Mb) and M33 (0.118 Mb region), we designed 12 primer sets semi-manually that flank the observed polymorphic SNPs using allele-specific to R haplotype and alternate allele specific to S haplotype. Among these 12 primer sets, two primer sets were from those contigs that anchored to chromosome 0 whereas rest 10 primer sets were from contigs that anchored to chromosome 11 (Table 13).

Table 13 Primers sets developed using trait-specific unique SNPs

Label	Primer name	Primers (5'----> 3')	Amplicon length	Position of contig in DM
KASP_1_ 7358	KP_FR_C4038	gaaggtgaccaagttcatgctttatacagatttcaagttcgagttcta	157 bp	Chr 0, 20816801 – 20818460 bp
	KP_FS_C4038	gaaggtcggagtcaacggattttatacagatttcaagttcgagttctg		
	KP_CR_C4038	agagcgtcacataaattgtgg		
KASP_2_ 7358	KP_RR_C4022	gaaggtgaccaagttcatgctcatacgtgtcacacttgaatatacag	114 bp	Chr 0, 20881875 – 20881465 bp
	KP_RS_C4022	gaaggtcggagtcaacggattcatacgtgtcacacttgaatatacaa		
	KP_CF_C4022	acttcgccagatacaatcatct		
KASP_3_ 7358	KP_RR_S22	gaaggtgaccaagttcatgctcatgcagttataagtcaggtgtaca	190 bp	Chr 11, 1064255-1063487 bp
	KP_RS_S22	gaaggtcggagtcaacggattcatgcagttataagtcaggtgtacg		
	KP_CF_S22	Ccctctccatttctgcactg		
KASP_4_ 7358	KP_FR_C4188	gaaggtgaccaagttcatgctgacatcccgaacctataaagttg	87 bp	Chr 11, 1150124-1152604 bp
	KP_FS_C4188	gaaggtcggagtcaacggattgacatcccgaacctataaagttt		
	KP_CR_C4188	Aatcgccggagcttttagtt		
KASP_5_ 7358	KP_RR_S18	gaaggtgaccaagttcatgctttgggacaccgactggaaa	171 bp	Chr 11, 1065929-1065101 bp
	KP_RS_S18	gaaggtcggagtcaacggatttgggacaccgactggaac		
	KP_CF_S18	ttttaaacggagggagtagatatgtt		
KASP_6_ 7358	KP_FR_S101	gaaggtgaccaagttcatgctggattcaaacctagattaagcatc	87 bp	Chr 11, 1259860-1261341 bp
	KP_FS_S101	gaaggtcggagtcaacggattggattcaaacctagattaagcatt		
	KP_CR_S101	Cgtgcttttgaatggtctatg		
KASP_7_ 7358	KP_FR_S258	gaaggtgaccaagttcatgctctgaagcagtcctgcagat	102 bp	Chr 11, 1186753-1191496 bp
	KP_FS_S258	Gaaggtcggagtcaacggattctgaagcagtcctgcagac		
	KP_CR_S258	tccttgaggagaaagtaagtgtg		
	KP_FR_C2880	gaaggtgaccaagttcatgctttctccacttagatctcacgttttt	51 bp	

KASP_8_7358	KP_FS_C2880	gaaggtcggagtcaacggattttctccacttagatctcacgttttc		Chr 11, 859508 – 858536 bp
	KP_CR_C2880	cgatatgtttcactgcaattgat		
KASP_9_7358	KP_FR_C3896	gaaggtgaccaagttcatgctagcccttccttccgcata	91 bp	Chr 11, 907676 – 906677 bp
	KP_FS_C3896	gaaggtcggagtcaacggattagcccttccttccgcatg		
	KP_CR_C3896	aaccatcactgcaagcgact		
KASP_10_7358	KP_RR_C3898	gaaggtgaccaagttcatgctcttgaaatctctaaccaggaatgc	76 bp	Chr 11, 962179-961178 bp
	KP_RS_C3898	gaaggtcggagtcaacggattcttgaaatctctaaccaggaatga		
	KP_CF_C3898	Ttcaatttgccggtcgag		
KASP_11_7358	KP_RR_C3940	gaaggtgaccaagttcatgcttgtaccaaacgatccttcaatg	87 bp	Chr 11, 802806-801757
	KP_RS_C3940	gaaggtcggagtcaacggatttgtaccaaacgatccttcaata		
	KP_CF_C3940	Tgtttacggggtgaaggttt		
KASP_12_7358	KP_RR_C3998	gaaggtgaccaagttcatgctccttatacttcctccacctacctat	156 bp	Chr 11, 848544-849232
	KP_RS_C3998	Gaaggtcggagtcaacggattccttatacttcctccacctacctaa		
	KP_CF_C3998	Atcctgtcaccactgagcttc		

Tail FAM (gaaggtgaccaagttcatgct) was added to the resistance allele-specific primer while tail HEX (gaaggtcggagtcaacggatt) was added to the susceptible allele-specific primer. RR = reverse and resistance-specific primer; RS = reverse and susceptible-specific primer; FR = forward and resistance-specific primer, FS = forward and susceptible-specific primer, CF = common forward primer, CR = common reverse primer; chr = chromosome.

3.7 Marker analysis and mapping

First, we checked on an agarose gel (2%) if the primers can amplify a product from a small number of DNA samples of resistant and susceptible genotypes. Next, we performed PCR using fluorescent labelling of FAM and HEX tails and determined fluorescence in FAM and HEX channels. PCR program was optimised by decreasing the annealing temperature from 56 to 50 ^{0}C. The primer pairs that performed according to expectation were selected for testing the entire population to see if clustering occurred.

Finally, the markers 8 (KASP_8_7358), maker 9 (KASP_8_7358), and marker 10 (KASP_8_7358) performed according to expectation. Using best melting temperature, marker 8 (55-56 ^{0}C), marker 9 (55 ^{0}C), and marker 10 (53 ^{0}C) were tested to DNA samples from a cap7358 population and small number of DNA samples of resistant and susceptible genotypes from Athlete x Queen Anne population. Interestingly, there were two clusters representing samples with each resistance allele and a susceptible allele for marker 8, 9 and 10. Athlete X Queen Anne population members and MQ (no template DNA) positioned in between these two clusters or more towards susceptible cluster for all three markers. Three susceptible genotypes such as Rpi05-7358-29, 7358-306, and Rpi05-7358-rec362 were clustered towards samples having resistance allele in all tested three markers. Again, Rpi05-7358-26 is associated with resistance phenotype, however, grouped more towards susceptible cluster (Appendix XIV). These four plants are most likely recombinants. We also found some level of contamination in assay most likely due to the use of MQ, which was not nuclease-free. The details of KASP genotyping for these 3 markers were depicted in Table 14.

Markers we developed were not clearly line up with the markers developed before. Therefore, based on most likely position to reference genome DM and recombinants observed, the genetic map was constructed. Marker 8, Marker 9 and Marker 10 are located on 0.8, 0.9 and 0.96 Mb region of chromosome 11 (Figure 12).

Table 14 KASP genotyping result for three most promising markers

Samples	Use status in sequencing	Phenotype	KASP_10_7358	KASP_9_7358	KASP_8_7358
CAP536-1	N	R	R	ND	R
CRC564-3	N	S	S	ND	ND
Rpi05-7358-5	Y	R	R	R	R
Rpi05-7358-9	Y	R	R	R	R
Rpi05-7358-10	Y	S	S	S	S
Rpi05-7358-11	Y	S	S	S	S
Rpi05-7358-12	N	R	R	R	R
Rpi05-7358-26	N	R	S	S	S
Rpi05-7358-29	Y	**S**	R	R	R
Rpi05-7358-47	Y	S	S	S	S
Rpi05-7358-rec355	N	S	S	S	S
Rpi05-7358-rec362	N	**S**	R	R	R
7358-148	Y	S	S	S	S
7358-213	Y	S	S	S	S
7358-232	N	NA	R	R	R
7358-275	N	NA	S	S	S
7358-276	Y	S	S	S	S
7358-280	N	S	S	S	S
7358-291	N	R	R	R	R
7358-301	Y	R	R	R	R
7358-305	N	NA	R	R	R
7358-306	Y	S	R	R	R
7358-321	Y	R	R	R	R
7358-328	Y	R	R	R	R
7358-344	Y	R	R	R	R
7358-350	N	NA	S	S	S
7358-355	Y	S	S	S	S
7358-360	Y	R	R	R	R
7358-362	N	S	**S**	R	**S**
7358-363	N	S	S	S	S
7358-3b	N	R	R	R	R
7358-3b20	N	NA	R	R	R
7358-S3	Y	S	S	S	S

Y = Yes; N = No; R = Resistant; S = Susceptible; NA = Not available; ND = Not determined

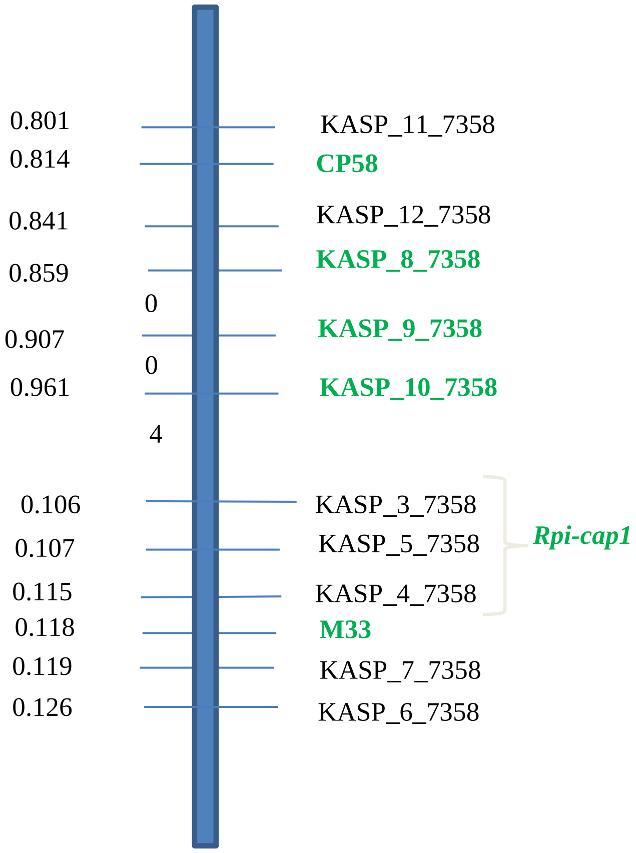


Figure 12 The genetic map of the resistance gene, *Rpi-cap1* on chromosome 11 of potato genome. The whole number on left side of the map shows the number of recombinants out of 26 progeny population having phenotypic information. The number with a decimal in left side indicates the physical position of the markers that are shown on the right side.

4 DISCUSSION

4.1 Bulked segregant analysis: effectiveness and optimization

Following an intraspecific crossing between *S. capsibaccatum* resistant and susceptible parent, the segregating progenies were assayed for two contrasting resistant and susceptible traits. The DNA samples were pooled for each trait followed by whole-genome sequencing. It was expected that the allele frequency for two pools or bulks should be roughly equal except the causal genomic region or loci, which, indeed exhibiting different allele frequency (Hart *et al.*, 2015; Li *et al.*, 2017). Computer algorithms aided bulked segregant analysis (BSA) on pooled DNA sequencing data help to identify target trait specific allelic variation of causal loci and thereby aid several benefits. For example, cost and time effectiveness, simplicity on use over the use of near-isogenic lines or other mapping populations, however, segregating population is needed to select resistance and susceptible plant (Giovannoni *et al.*, 1991; Michelmore *et al.*, 1991; Warburton *et al.*, 2010; Terauchi *et al.*, 2015; Zou *et al.*, 2016). This approach is even tolerable to accidental phenotyping inaccuracies (Schneeberger *et al.*, 2009). In the present study, PCR based flanking markers had been already reported which were linked to the target gene, *i.e. Rpi-cap1* (Verzaux *et al.*, 2012). The marker and gene distances in genetic mapping still need to be narrowed (high-resolution genetic mapping) to provide opportunity in introgression breeding.

Five different bulked segregant k-mers analysis (BSKA) approaches were used to derive a k-mers list of resistance specific haplotype (Table 3). BR-specific k-mers list with different coverage threshold was mapped to reference DM to find the putative position of resistance locus on potato chromosome. All approaches produced the result in the same line: the higher numbers of k-mers mapped to 0-2 Mb region of chromosome 11, the region where the gene is located. This result proved the haplotype concept that we hypothesized during construction of five different BSKA approaches (Table 3).

Charlotte Prodhomme (unpublished data) used BSA approach I which retained the BR specific k-mers shared with PR but not present in BS. The present study showed this approach kept the least k-mers volume. Due to the use of intersection function between BR and PR during the first step of the analysis, this approach kept only those k-mers which have less coverage depth among PR and BR. (Table 3). Most likely, this approach does not include k-mers which are underrepresented in the NGS data of PR and BR. However, BSA approach I might be subjected to least errors. When mapping, the SNR combined with a count of mapped k-mers on the target region of the genome give more information than only number of k-mers. The BSKA approach V which includes only BR and BS during analysis (BR-BS) recorded the highest SNR for the target region of chromosome 11 than any others. This showed the absence of PR does not affect on getting signal in the target region of interest for current research. It was also supported by BSA approach III which retained k-mers in PR but not in BS (PR-BS), where SNR in the target region was very low. We don’t have much information to explain the reason but maybe sampling during sequencing in the present study is responsible. This approach, therefore, might miss some important alleles and BSA approach V (BR-BS) might be a potential approach for downstream analysis. Even the cost of the project would be highly reduced if this approach detects the trait associated genomic region and allelic variation with same statistical power. There are several literature that support the use of NGS data on two bulks but not the parent during BSA to determine the trait-specific region of interest (Terauchi *et al.*, 2015). In the present study, the BSA approach IV which lists all k-mers found in PR and BR but not in BS retained the highest number of k-mers that mapped to an entire region of the genome and the target region. Those k-mers with less coverage but specific to resistance bulk will be retained due to higher coverage (use of union function). Theoretically, BSA approach IV always kept higher R haplotype frequency due to the addition of R haplotype frequency present in PR and BR (Table 15). Therefore, we selected BSKA approach IV as a suitable approach for downstream analysis. We assumed same rule also applied to derive k-mers list specific to S haplotype.

If candidate gene region is already known (as in present study), the use of NGS data on PR (or BR) and BS is sufficient to meet our objective, *i.e.* allele mining and mapping the causal *R* gene, however, the power of detection of an allele may vary. Theoretically, there are four haplotypes R, r, r' and r'' and the expected R haplotype frequency is not always same for all BSKA approaches. For example, R haplotype frequency in BSKA approach that uses only resistant parent and susceptible bulk retained lesser R haplotype frequency than in another approach that utilizes PR, BR and BS (Refer Table 3 and Table 15). Besides, the higher SNR combined with the increased number of mapping k-mers to the target region of the genome may increase the power of further downstream analysis (read retrieval, de-novo assembly, mapping, SNP mining etc.).

Table 15 Frequency of each haplotype present in resistant parent, susceptible parent, resistant bulk and susceptible bulk

Haplotype	Frequency			
	PR	PS*	BR	BS
R	0.5	0	0.5	0
R	0.5	0	0	0.5
r'	0	0.5	0.25	0.25
r"	0	0.5	0.25	0.25

PR = resistant parent; PS = susceptible parent; BR = resistant bulk; BS = susceptible bulk; Refer Appendix VI to have clear overview regarding how haplotype was retained in each bulk and parent. * represents the NGS data on whole genome sequencing was not available for the present study

Besides expected region on chromosome 11, we also found k-mers peak on chromosome 12. This peak must be an artefact as the presence or absence of the *Rpi-cap1* gene in the bulk individuals was tested using molecular markers. So, these bulks were built to enrich for a genomic region and not for multi-locus traits. The artificial chromosome 12 region must be a haplotype from the PR as all BSKA approaches that we derived inflicted with artefacts or the absence of one of the resistance partner (PR or BR) does not help to resolve the problem. Most likely the small size of bulks (11 plants on each bulk) used in the present study is responsible for this artefact (or the selection of 22 plants does not remove the artefacts). To detect true genetic position or QTL, one should choose a big size of bulk (Magwene *et al.*, 2011; Zou *et al.*, 2016). In such condition, selective genotyping of individual members of pooled bulk provides a cost-efficient genetic mapping that most likely represents the entire population (Sun *et al.*, 2010). The ideal situation is not always possible, therefore we doubt effectiveness on use of BSA where NGS data on whole genome sequencing were produced from small bulk size. Still, the validation is time-consuming and the limited recombination in few individuals of bulk hampers the gene mapping (Li *et al.*, 2018). Fortunately, individuals from bulks in the present study were pre-selected for having a recombination in target genomic region. This might be the reason of getting a high and narrow peak in chromosome 11, 0-2 Mb region which otherwise would be a bell-shaped curve in bulks derived from a normal F1 population. Still, the peak region is broader in the target region of chromosome 11 than expected. Under such situation one can expect the reduced value of the signal (signal may be diluted), thus producing lower SNR, however, that might apply to all the approaches we developed. As many plants in the bulks are recombinants, one might expect to identify narrow candidate gene region with more power when reducing the bin size from 1 Mb (as in present study) to 100 Kb or 50 Kb. Under such situation, we may get the increased value of the signal. Also, the minimum mapping quality criterion set during analysis may influence on SNR. Being affected by several factors, use of SNR should be done with caution.

4.2 Alignment, assembly and variant calling

The decreasing price for sequencing and availability of more and more reliable computer algorithms has increased the efficiency of big data analysis. Several algorithms have been developed for read alignment or mapping. We used Burrows-Wheeler Aligner (BWA) for short read alignment as this package is faster and allow to use other pipelines on its alignment output file, for example use of SAMtools to select, sort, merge alignment region and even allow to call variants (Li and Durbin *et al.*, 2009; Li and Durbin *et al.*, 2010). Based on the specificity of the programs that BWA package have, we used BWA-backtrack and BWA-mem for k-mers mapping and read mapping, respectively. After selecting bulk-specific k-mers and mapping them to potato reference genome DM, only single copy number k-mers in the target region of interest, *i.e.* chromosome 11, 0-2 Mb region and chromosome 0, 20800 Kb to 20950 Kb region were selected called haplotype-specific k-mers. The coverage of haplotype was observed according to our expectation (half of the sum of BR and PR). The SNPs density was calculated and was found very low (0.55 SNP/Kb for R haplotype and 0.76 SNP/Kb for S haplotype relative to DM) in the target region of interest implying contiguous assembly is unlikely (Refer Table 6). Apparently, under such condition, the k-mers that do not well matched or mapped to reference genome due to less similarity would get removed and we may miss some important SNPs found to resistance haplotype in relative to susceptible haplotype. Therefore, *de-novo* assembly was done. Still, the high copy number sequence impairs proper *de-novo* assembly of sequence (Jupe *et al.*, 2013; Andolfo *et al.*, 2014). So, we retrieved the single copy number haplotype-specific original sequence read using above-produced haplotype-specific k-mers list. The R and S haplotype-specific reads mapping to a target region of the reference genome DM were *de-novo* assembled to create haplotype-specific contigs. Next mapping of the haplotype-specific reads to the R contigs allowed to determine haplotype-specific variants. The variants filtering is still tricky (Magwene *et al.*, 2011). Among the variants, SNPs are most common and powerful biallelic form of potato genetic variation (Potato Genome Consortium, 2011) which can be readily used to make trait-specific markers. The SNPs with lower coverage may be caused by sampling error, leading many false positives, however, SNPs with higher coverage can lead to a selection of multi-copy markers. To increase the probability of capturing good SNPs, present study relied on selecting k-mers coverage. This allowed us to select only single copy reads. So, FreeBayes was fed with single copy reads only. However, if the process selects multi-copy reads by chance than variant detector tools (FreeBayes) we used in the present study can't filter out (FreeBayes can't fix the maximum threshold efficiently). The solution might be taking care during visualization on IGV after getting variants information. Still, this is manual and possibly prone to error.

The FreeBayes algorithm detects more than 600 additional SNPs in R and S haplotype relative to DM than we expected (Refer Table 6 and Figure 11) in the target region of interest. The result was found to be obvious as SNPs were not equally distributed over the whole genome. Some regions of the genome were enriched with SNPs whereas others not (Appendix XV; Appendix XVI). Again, the use of different pre-filtering and post-filtering setting greatly influenced on haplotype-based variants mining (Garrison and Marth, 2012). We found a lower number of SNPs (2,315) in the target region of S haplotype when it aligned to R contigs than reference genome DM. In the present study, the *de-novo* assembly produced a number of contigs with smaller to medium sizes. Also, scaffolding has not been performed in the maximal way (Refer Table 9). Due to the smaller size of contigs, there was the higher chance of mapping paired reads on another contig when haplotype-specific reads mapped to *de-novo* assembled contigs (Refer Table 11 specifically third column) and this reduces the mapping quality of reads. Computer algorithm (FreeBayes) does not call SNPs from lower mapping quality reads (Refer Appendix III) and that might be the reason for having reduced number of SNPs in the final list.

Still, a number of SNPs, both estimated and real, in susceptible haplotype is higher than the resistance haplotype (Refer Table 6 and Figure 11). It can be explained from evolutionary perspectives. First, R haplotype may underwent duplication and retained in the same cluster in the plant genome. Panchy et al. (2016) reported that the duplication is the normal way of evolution of the genes that are responsible for adaptation, thus allowing phenotypic novelty in the plant including disease resistance. Most likely, the best-known example is the *NBS-LRR* gene family (Leister, 2004). Indeed, duplication might result in multicopy k-mers which eventually removed during the analysis and the final list retained only single copy k-mers. This logic is also supported by the data regarding a number of unique k-mers specific to R and S haplotype, where S haplotype recorded a higher number of k-mers list than R haplotype (Refer Table 7). One can expect a higher number of SNPs if the particular haplotype contains a higher number of k-mers or reads.

For primer design, we selected only those SNPs that were found in S haplotype relative to *de-novo* assembled R contigs due to easiness on use. We designed markers from only those contigs that anchored to the target region (chromosome 11, 0-2 Mb region and chromosome 0, 20800 Kb to 20950 Kb region) of reference genome DM. Those contigs that unanchored to the region of interest on chromosome 11, DM was not used. However, there may be a chance that the blast setting used in the present study may not let contigs to anchor in any region of the genome and still contains the useful SNPs. Maybe there is less similarity between *de-novo* assembled contigs and DM but they lie in the target region of the genome and associated with the resistance trait. Still, the present study did not rely on selecting contigs that retained *NB-LRR* gene-like sequences based on homology of the sequence.

4.3 Marker development and validation in segregating population

As discussed before, BSA approach efficiently can determine the genetic variations (polymorphisms) between the R haplotype and S haplotype for development of PCR marker. In the present study, we validated three markers that co-segregate with resistance trait on cap7358 population. Under given condition (Refer Table 3) one can expect heterozygous allele on a PR, resistance allele only or heterozygous in resistance progeny members and alternate susceptible allele on susceptible progeny members on individual KASP genotyping result. Theoretically, the susceptible parent would not contain allele-specific to susceptible haplotype as described in Appendix VI unless and until the same susceptible-specific haplotype is present in both resistance and susceptible parent. In line with expectation, most of the resistant bulk progeny members and susceptible bulk progeny members were grouped to resistance and susceptible cluster with appropriate fluorescence tag, respectively. But in contrast to expectation, the PR found to contain only resistance allele for marker 8 and marker 10 but not in marker 9. The susceptible allele was present in a susceptible parent for marker 10, which is not in line with our expectations. This might indicate the allele we used for marker development is tri-allelic SNP. Maybe the selected susceptible specific haplotype might not be coming from the PR. For maker 8 and marker 9, genotyping result not showed any clear indication of having susceptible specific allele in susceptible parent which was according to our expectation, however, for PR, again it contains only resistance allele or the allele was not determined. The absence of a susceptible allele in PR may indicate the primers specific to susceptible allele may have lower affinity than primers specific to resistance allele to PCR or the SNPs might be tri-allelic again. Intriguingly, we found 4 recombinants, which is a high figure, however, it seems logical as the selection of individuals from bulks in the present study were enriched for recombination between M33 and Cp58 markers.

Using the information on most likely marker position relative to reference genome DM, a genetic map was built manually. Three validated polymorphic KASP markers from the present study were found to be located closer to *Rpi-cap1* gene than the CAPS markers suggested by Verzaux et al. (2012). Even the current markers were developed using diploid wild germplasm, the assay nature of KASP marker (quantitative assay) could provide a more valid assay to determine each zygosity level (simplex, duplex, triplex etc) according to calculation of FAM and HEX signal ratios, when testing in tetraploid potato (Uitdewilligen *et al.,* 2015). Even KASP genotyping is a more flexible solution over CAPS genotyping used before as useful genetic variation not need to have restriction enzyme recognition site (Patterson *et al.,* 2017). The primers used in the present KASP assay, however, do not include the stretches of sequences that contain multiple SNPs. The use of single SNPs during primer construction is sufficient for genotyping of sample DNA, however using multiple SNPs improved the reliability and robustness of genotyping (Patterson *et al.*, 2017).

We found three markers that co-segregate with *Rpi-cap1*. The markers test on Athlete x Queen Anne population showed Athlete does not contain the resistance gene, *Rpi-cap1*. It might be the case that another *NB-LRR* gene in Athlete recognized *Avr-cap1* effector. The database showed one of the Athlete parents, Miriam provide a medium level of resistance, however, another parent AR 99-263-5 has not been characterized and the details information is not readily available (Berlo *et al.*, 2007; http://10.73.177.202/potatopedigree/). The collection of AR 99-263-5 accession from gene bank (if available) and phenotypic characterization might forecast basic characteristics features of AR 99-263-5. The progeny population in Athlete and Queen Anne crossing is less than the present study. One might expect serious artefacts if gene mining in Athlete would be based on same BSA approach.

4.4 Applicability of KASP assay in small companies and developing countries

Wageningen University and research have been doing KASP Genotyping using KASP Mastermix produced and distributed by LGC genomics. With their monopoly market, they offer KASP Master Mix at relatively higher prices, therefore there might be problems on the utilization of developed KASP markers for small institute or company. Generally, screenings of markers need to test a big number of putative SNPs and hundreds of individuals, which make the KASP genotyping expensive. This regard, self-made Amplifluor-like SNP system may provide a better choice over KASP in terms of cost and flexibility (Jatayev *et al.*, 2017). Moreover, the CAPS marker has a wider use for small to medium-scale experiments and can be run in the very basic laboratory. This marker type is again more applicable if the genetic region is highly polymorphic (Shavrukov, 2016)

5 CONCLUSION AND RECOMMENDATIONS

The 1:1 segregation conferred by an intraspecific cross between *S. capsibaccatum* resistance and susceptible accession showed single dominant resistance gene, *Rpi-cap1* is responsible for resistance trait. Few *Rpi-cap1* flanking markers from conventional techniques were already described. The present study relied mainly on bulked segregant k-mers analysis (BSKA). The BSKA was done to select single copy R and S haplotype-specific k-mers. The target region for respective haplotype was determined for both chromosome 11 (0-2 Mb bin region) and chromosome 0 (20800-20950 Kb bin region). We confirmed the resistance locus and for the first time, haplotype-specific SNPs were identified and listed successfully. We reported 3261 and 4432 unique SNPs in R haplotype and S haplotype relative to potato reference genome DM, respectively. There were 2315 SNPs specific to S haplotype relative to *de-novo* assembled R contigs. Using SNPs in S haplotype relative to R contigs that anchored to the target region of potato genome, we developed 12 haplotype-specific KASP primer sets from putative gene region. Three KASP markers were verified as polymorphic which were closer to the *Rpi-cap1* gene than before. These polymorphic markers might be interesting markers for potato breeders. Again, the multiple markers can be used in future to increase the robustness of detection. However, before utilization of markers developed in current study into introgression resistance breeding, fine mapping followed by map-based cloning and functional study (eg. gene or RNA silencing, RNA interference, etc) of a cloned gene could be a better choice.

Breeders are interested in markers that flank the gene in both sides for screening the germplasm, however, the present study verified the KASP makers that flank the gene from one of the side. We narrowed down the location of *Rpi-cap1* gene from ~0.35 Mb to ~0.2 Mb (Figure 12). We found 4 recombinants. First, we would recommend re-phenotyping of recombinant genotypes such as Rpi05-7358-29, 7358-306, and Rpi05-7358-rec362 and Rpi05-7358-26 to determine whether they are real recombinant or phenotyping error. To get markers from both ends of a gene, it is recommended to validate markers that positioned on the northern side of chromosome 11, for example, 0.96 Mb to 0.116 Mb region. Still, there are hundreds of potential haplotype-specific SNPs relative to reference genome DM and *de-novo* assembled contigs. I would recommend to construct and validate regular interval single copy KASP (including self-made Amplifluor-like SNP system) and CAPS marker utilizing haplotype-specific SNPs and Indels as well. We would strongly recommend using nuclease-free MQ to get true fluorescent read during KASP genotyping. Moreover, one can explore the idea to determine the resistance specific contigs closest to the gene region among the list of *de-novo* contigs. The homology of sequence from *de-novo* contigs could be checked for *NB-LRR* gene. Also, the haplotype-specific k-mers and read list we produced in the present study could be useful to develop the markers for other resistance genes which are located in 0-2 Mb region of potato chromosome 11.

The present study reported no *Rpi-cap1* gene in Athlete x Queen Anne tetraploid crossing population. It has been found that one of the parents of Athlete (AR 99-263-5) has not been characterized, therefore it would be better to produce information on a resistant parent of Athlete through characterization of AR 99-263-5 accession. If there is difficulty in getting AR 99-263-5 accession, BSA approach could be done but with precautions because of the small bulk size we have for Athlete x Queen Anne population. We would recommend increasing population size and sequencing depth of whole genome of bulks and parent(s) of Athlete x Queen Anne population to determine homozygous variants. If this condition could be met, one should rely on another alternative strategy called resistance gene enrichment sequencing (RenSeq). The short Renseq Illumina reads produces *NB-LRR* contig where a researcher can align reads from parent and bulks and thereby SNPs calling is possible within the member of the gene family (Jupe *et al.*, 2013). The Solanum bait library used by Jupe et al. (2012) and Jupe et al. (2013) again could be used to capture the NLR. This technique could help to verify old *NB-LRR* or may annotate new gene family. The improvement has been made in RenSeq in regard to using single molecule real time sequencing (SMRT) over short parallel sequencing. The use of reads generated by SMRT RenSeq helps to determine the full sequence of a gene (Witek *et al.,* 2016). Again, gene polymorphisms could be checked in parents and in segregating progeny.

It may be difficult to get higher coverage for the larger genome for every region of the chromosome, for example, potato. From above sections, it is clear that the small size of bulk may contain artefacts during BSA, therefore, one should always try to get the higher size of segregating progeny and size of samples that constitute the bulks. Again, the parallel sequencing most likely results gaps between contigs, however, the Pacbio read fill the gap between contigs, thus resulting in longer contigs (Oppelaar, 2017). In order to get true reference contig, the hybrid assembly that uses both Illumina and Pacbio read (example, Abyss) would be a better choice than the sole use of NGS parallel sequencing

The present research used SOAPdenovo2 for *de-novo* assembly of reads, which is time-consuming to configure the input file and later select the appropriate k-mers size. This regard, SPADES might be better computer algorithm which automatically selects the k-mers size and working pipeline is easy (based on personal experience). Still, FreeBayes would be best to call the variants as it can consider the ploidy level of the genome. To select the accurate SNPs and less false positives, future research should focus on getting sufficient sequence coverage and more stringent analysis pipeline.

In summary, the present study showed the potential of applying BSA combined with de-novo assembly and haplotype-based variant calling pipelines for the identification of causal genomic locus and haplotype-specific allelic variations associated with trait specific bulk.

6 LITERATURE CITED

Andolfo, G., Jupe, F., Witek, K., Etherington, G. J., Ercolano, M. R., & Jones, J. D. (2014). Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC plant biology*, *14*(1), 120.

Birch, P. R., Bryan, G., Fenton, B., Gilroy, E. M., Hein, I., Jones, J. T., Prashar, A., Taylor, M.A., Torrance, L., & Toth, I. K. (2012). Crops that feed the world 8: Potato: are the trends of increased global production sustainable?. *Food Security*, *4*(4), 477-508.

Bradshaw, J. E. (2009). Potato breeding at the Scottish plant breeding station and the Scottish Crop Research Institute: 1920–2008. *Potato research*, *52*(2), 141-172.

De Buck, A. J., Van Rijn, I., Roling, N. G., & Wossink, G. A. A. (2001). Farmers' reasons for changing or not changing to more sustainable practices: an exploratory study of arable farming in the Netherlands. *The Journal of Agricultural Education and Extension*, *7*(3), 153-166.

Drenth, A., Turkensteen, L. J., & Govers, F. (1993). The occurrence of the A2 mating type of *Phytophthora infestans* in the Netherlands; significance and consequences. *European Journal of Plant Pathology*, *99*, 57-67.

Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., & Nguyen, N. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, *21*(12), 2224-2241.

Ertiro, B. T., Ogugo, V., Worku, M., Das, B., Olsen, M., Labuschagne, M., & Semagn, K. (2015). Comparison of Kompetitive Allele Specific PCR (KASP) and genotyping by sequencing (GBS) for quality control analysis in maize. *BMC genomics*, *16*(1), 908.

FAOSTAT. (2016). Region/World/Production Quantity/Crops from pick lists". UN Food and Agriculture Organization, Statistics Division. Available on: http://www.fao.org/faostat/en/#data/QC [retrieved on 24 July 2017].

Fiers, M., Edel-Hermann, V., Chatot, C., Le Hingrat, Y., Alabouvette, C., & Steinberg, C. (2012). Potato soil-borne diseases. A review. *Agronomy for Sustainable Development*, *32*(1), 93-132.

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annual review of phytopathology*, *9*(1), 275-296.

Fry, W. (2008). *Phytophthora infestans*: the plant (and R gene) destroyer. *Molecular plant pathology*, *9*(3), 385-402.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Gebhardt, C., & Valkonen, J. P. (2001). Organization of genes controlling disease resistance in the potato genome. *Annual review of phytopathology*, *39*(1), 79-102.

Giovannoni, J. J., Wing, R. A., Ganal, M. W., & Tanksley, S. D. (1991). Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. *Nucleic Acids Research*, *19*(23), 6553-6568.

Govers, F., Drenth, A., & Pieterse, C. M. J. (1997). The potato late blight pathogen *Phytophthora infestans* and other pathogenic oomycota. The Mycota: plant relationships, pp. 17 – 36

Hart, J. P., & Griffiths, P. D. (2015). Genotyping-by-sequencing enabled mapping and marker development for the potyvirus resistance allele in common bean. *The Plant Genome*, *8*(1).

Hawkes, J. G., & Hjerting, J. P. (1989). The potatoes of Bolivia: their breeding value and resistance in the potato genome. Annual Review of Phytopathology 39:79-102.

Hein, I., Birch, P. R., Danan, S., Lefebvre, V., Odeny, D. A., Gebhardt, C., Trognitz, F., & Bryan, G. J. (2009). Progress in mapping and cloning qualitative and quantitative resistance against *Phytophthora infestans* in potato and its wild relatives. *Potato Research*, *52*(3), 215-227.

Hogenhout, S. A., Van der Hoorn, R. A., Terauchi, R., & Kamoun, S. (2009). Emerging concepts in effector biology of plant-associated organisms. *Molecular plant-microbe interactions*, *22*(2), 115-122.

Hwang, Y. T., Wijekoon, C., Kalischuk, M., Johnson, D., Howard, R., Prüfer, D., & Kawchuk, L. (2014). Evolution and management of the Irish potato famine pathogen *Phytophthora infestans* in Canada and the United States. *American journal of potato research*, *91*(6), 579-593.

Jacobs, M. M., Vosman, B., Vleeshouwers, V. G., Visser, R. G., Henken, B., & van den Berg, R. G. (2010). A novel approach to locate *Phytophthora infestans* resistance genes on the potato genetic map. *Theoretical and applied genetics*, *120*(4), 785-796.

Jatayev, S., Kurishbayev, A., Zotova, L., Khasanova, G., Serikbay, D., Zhubatkanov, A., Botayeva, M., Zhumalin, A., Turbekova, A., Soole, K., Langridge P., & Langridge, P. (2017). Advantages of Amplifluor-like SNP markers over KASP in plant genotyping. *BMC plant biology*, *17*(2), 254.

Jones, J. D., & Dangl, J. L. (2006). The plant immune system. *Nature*, *444*(7117), 323.

Jupe, F., Pritchard, L., Etherington, G. J., MacKenzie, K., Cock, P. J., Wright, F., Sharma, S. K., Bolser, D., Bryan, G. J., Jones J. D. G. & Hein, I. (2012). Identification and localisation of the *NB-LRR* gene family within the potato genome. *BMC genomics*, *13*(1), 75.

Jupe, F., Witek, K., Verweij, W., Śliwka, J., Pritchard, L., Etherington, G. J., Maclean D., Cock, P. J., Leggett, R. M., Bryan, G. J., Cardle, L., Hein, I., & Jones J. D. G. (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*(3), 530-544.

Kaplinski, L., Lepamets, M., & Remm, M. (2015). GenomeTester4: a toolkit for performing basic set operations-union, intersection and complement on k-mer lists. *Gigascience*, *4*(1), 58.

Kim, H. J., Lee, H. R., Jo, K. R., Mortazavian, S. M., Huigen, D. J., Evenhuis, B., Kessel, G., Visser, R.G.F., Jacobsen E. & Vossen, J. H. (2012). Broad spectrum late blight resistance in potato differential set plants MaR8 and MaR9 is conferred by multiple stacked R genes. *Theoretical and applied genetics*, *124*(5), 923-935.

Kromann, P., Pradel, W., Cole, D., Taipe, A., & Forbes, G. A. (2011). Use of the environmental impact quotient to estimate health and environmental impacts of pesticide usage in Peruvian and Ecuadorian potato production. *Journal of Environmental Protection*, *2*(5), 581.

Kumpatla, S. P., Buyyarapu, R., Abdurakhmonov, I. Y., & Mammadov, J. A. (2012). Genomics-assisted plant breeding in the 21st century: technological advances and progress. In *Plant breeding*. InTech.

Leister, D. (2004). Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in genetics*, *20*(3), 116-122.

Li P., Du C., Zhang Y., Yin S., Zhang E., Fang H., Lin D., Xu C., & Yang Z. (2018) Combined bulked segregant sequencing and traditional linkage analysis for identification of candidate gene for purple leaf sheath in maize. *PLoS ONE 13*(1): e0190670.

Li, B., Zhao, Y., Zhu, Q., Zhang, Z., Fan, C., Amanullah, S., Gao, P., & Luan, F. (2017). Mapping of powdery mildew resistance genes in melon (Cucumis melo L.) by bulked segregant analysis. *Scientia Horticulturae*, *220*, 160-167.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589-595.

Li, Y., van der Lee, T. A. J., Evenhuis, A., van den Bosch, G. B. M., van Bekkum, P. J., Förch, M. G., Gent-Pelzer, M. P. E., Raaij, H. M. G., Jacobsen, E., Huang, S. W., Govers, F., Vleeshouwers, V. G. A. A. & Kessel, G. J. T. (2012a). Population dynamics of *Phytophthora infestans* in the Netherlands reveals expansion and spread of dominant clonal lineages and virulence in sexual offspring. *G3: Genes, Genomes, Genetics*, *2*(12), 1529-1540.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Yang, B. & Fan, W. (2012b). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, *11*(1), 25-37.

Lozano, R., Ponce, O., Ramirez, M., Mostajo, N., & Orjeda, G. (2012). Genome-wide identification and mapping of NBS-encoding resistance genes in *Solanum tuberosum* group phureja. *PLoS One*, *7*(4), e34775.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., & Tang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*(1), 18.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., & Tang, J. (2015). Erratum: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *4*(1), 30.

Magwene, P. M., Willis, J. H., & Kelly, J. K. (2011). The statistics of bulk segregant analysis using next generation sequencing. *PLoS computational biology*, *7*(11), e1002255.

McDonald, B. A., & Linde, C. (2002). The population genetics of plant pathogens and breeding strategies for durable resistance. *Euphytica*, *124*(2), 163-180.

McHale, L., Tan, X., Koehl, P., & Michelmore, R. W. (2006). Plant NBS-LRR proteins: adaptable guards. *Genome biology*, *7*(4), 212.

Michelmore, R. W., & Meyers, B. C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome research*, *8*(11), 1113-1130.

Michelmore, R. W., Paran, I., & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the national academy of sciences*, *88*(21), 9828-9832.

Neelam, K., Brown-Guedira, G., & Huang, L. (2013). Development and validation of a breeder-friendly KASPar marker for wheat leaf rust resistance locus Lr21. *Molecular breeding*, *31*(1), 233-237.

Oppelaar D. (2017). End presentation on citrus *Colletotrichum* fungus (21 November, 2017) at Wageningen University and Research, The Netherlands. Internship at University of Florida, USA.

Panchy, N., Lehti-Shiu, M., & Shiu, S. H. (2016). Evolution of gene duplication in plants. *Plant physiology*, *171*(4), 2294-2316.

Patterson, E. L., Fleming, M. B., Kessler, K. C., Nissen, S. J., & Gaines, T. A. (2017). A KASP genotyping method to identify northern watermilfoil, Eurasian Watermilfoil, and their interspecific hybrids. *Frontiers in plant science*, *8*.

Potato Genome Sequencing Consortium. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, *475*(7355), 189-195.

Potts, M. J. (1990). Influence of intercropping in warm climates on pests and diseases of potato, with special reference to their control. *Field Crops Research*, *25*(1-2), 133-144.

Quarrie, S. A., Lazić-Jančić, V., Kovačević, D., Steed, A., & Pekić, S. (1999). Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. *Journal of experimental botany*, *50*(337), 1299-1306.

Rauscher, G. M., Smart, C. D., Simko, I., Bonierbale, M., Mayton, H., Greenland, A., & Fry, W. E. (2006). Characterization and mapping of *RPi-ber*, a novel potato late blight resistance gene from *Solanum berthaultii*. *TAG Theoretical and Applied Genetics*, *112*(4), 674-687.

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome research*, *20*(9), 1165-1173.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jorgensen, J., Weigel, D. & Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature methods*, *6*(8), 550-551.

Schouten, H. J., & Jacobsen, E. (2008). Cisgenesis and intragenesis, sisters in innovative plant breeding. *Trends in plant science*, *13*(6), 260-261.

Semagn, K., Babu, R., Hearne, S., & Olsen, M. (2014). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Molecular Breeding*, *33*(1), 1-14.

Shavrukov, Y. (2016). Comparison of SNP and CAPS markers application in genetic research in wheat and barley. *BMC plant biology*, *16*(1), 11.

Staskawicz, B. J., Ausubel, F. M., Baker, B. J., Ellis, J. G., & Jones, J. D. (1995). Molecular genetics of plant disease resistance. *Science*, *268*(5211), 661.

Sun, Y., Wang, J., Crouch, J. H., & Xu, Y. (2010). Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. *Molecular Breeding*, *26*(3), 493-511.

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S., & Terauchi, R. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *The Plant Journal*, *74*(1), 174-183.

Tamm, L., Smit, A. B., Hospers, M., Janssens, S. R. M., Buurma, J. S., Molgaard, J. P., & Bertand, C. (2004). Assessment of the socio-economic impact of late blight and state-of-the-art management in European organic potato production systems.

Tan, M. A., Hutten, R. C., Celis, C., Park, T. H., Niks, R. E., Visser, R. G., & van Eck, H. J. (2008). The *RPi-mcd1* locus from *Solanum microdontum* involved in resistance to *Phytophthora infestans*, causing a delay in infection, maps on potato chromosome 4 in a cluster of *NBS-LRR* genes. *Molecular Plant-Microbe Interactions*, *21*(7), 909-918.

Terauchi, R., Abe, A., Takagi, H., Tamiru, M., Fekih, R., Natsume, S., Yaegashi, H., Kosugi, S., Kanzaki, H., Matsumura, H., Saitoh, H., Yoshida, K., Cano, L., & Kamoun, S. (2015). Whole genome sequencing to identify genes and QTL in rice. In *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches* (pp. 33-42). Springer International Publishing.

Tiwari, J. K., Siddappa, S., Singh, B. P., Kaushik, S. K., Chakrabarti, S. K., Bhardwaj, V., & Chandel, P. (2013). Molecular markers for late blight resistance breeding of potato: an update. *Plant Breeding*, *132*(3), 237-245.

Tsedaley, B. (2014). Late blight of potato (*Phytophthora infestans*) biology, economic importance and its management approaches. *Journal of Biology, Agriculture and Healthcare, 4*(25), 215-225.

Uitdewilligen, J. G., Wolters, A. M. A., Bjorn, B., Borm, T. J., Visser, R. G., & van Eck, H. J. (2015). Correction: A Next-Generation Sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PloS one*, *10*(10), e0141940.

Van Berloo, R., Hutten, R. C. B., Van Eck, H. J., & Visser, R. G. F. (2007). An online potato pedigree database resource. *Potato research, 50*(1), 45-57.

Van der Vossen, E. A., Gros, J., Sikkema, A., Muskens, M., Wouters, D., Wolters, P., Pereira, A. & Allefs, S. (2005). The Rpi-blb2 gene from *Solanum bulbocastanum* is a Mi-1 gene homolog conferring broad-spectrum late blight resistance in potato. *The Plant Journal, 44*(2), 208-222.

Van Der Vossen, E., Sikkema, A., Hekkert, B. T. L., Gros, J., Stevens, P., Muskens, M., Wouters, D., Pereira, A., Stiekema, W., & Allefs, S. (2003). An ancient R gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to *Phytophthora infestans* in cultivated potato and tomato. *The Plant Journal*, *36*(6), 867-882.

Verzaux , E. (2010) Resistance and susceptibility to late blight in Solanum: gene mapping, cloning and stacking Thesis, Wageningen University, Wageningen, NL. ISBN 978-90-8585-631-3.

Verzaux, E., van Arkel, G., Vleeshouwers, V. G., van der Vossen, E. A., Niks, R. E., Jacobsen, E., Vossen, J., & Visser, R. G. (2012). High-resolution mapping of two broad-spectrum late blight resistance genes from two wild species of the *Solanum circaeifolium* group. *Potato research*, *55*(2), 109-123.

Vleeshouwers, V. G., Raffaele, S., Vossen, J. H., Champouret, N., Oliva, R., Segretin, M. E., Rietman, H., Cano, L. M., Lokossou, A., Kessel, G., Pel, M. A., & Kamoun, S. (2011). Understanding and exploiting late blight resistance in the age of effectors. *Annual review of phytopathology*, *49*, 507-531.

Warburton, M. L., Setimela, P., Franco, J., Cordova, H., Pixley, K., Bänziger, M., Dreisigacker, S., Bedoya, C., & MacRobert, J. (2010). Toward a cost-effective fingerprinting methodology to distinguish maize open-pollinated varieties. Crop science, 50(2), 467-477.

Watanabe, K. (2015). Potato genetics, genomics, and applications. *Breeding science*, *65*(1), 53-68.

Witek, K., Jupe, F., Witek, A. I., Baker, D., Clark, M. D., & Jones, J. D. (2016). Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. *Nature biotechnology*, *34*(6), 656-660.

Zou, C., Wang, P., & Xu, Y. (2016). Bulked sample analysis in genetics, genomics and crop improvement. *Plant biotechnology journal*, *14*(10), 1941-1955.

APPENDICES

Appendix I DNA from cap7358 population for whole genome sequencing

Sample ID	Phenotype	Nano ng/ul	260/280	260/230
CAP536-1	R	32.18	1.48	0.82
CRC564-3	S	26.91	1.59	0.90
7358-3b	R	15.36	1.50	0.80
Rpi05-7358-5	R	16.97	1.66	0.72
Rpi05-7358-9	R	16.56	1.80	1.00
7358-301	R	12.33	1.83	0.59
7358-328	R	12.53	1.70	0.73
7358-344	R	13.02	1.66	0.70
7358-219	R	15.30	1.79	0.77
Rpi05-7358-12	R	12.70	1.61	0.77
7358-360	R	25.54	1.73	1.04
7358-322	R	13.27	1.63	0.73
7358-321	R	11.14	1.69	0.61
7358-S1	S	13.87	2.06	0.80
Rpi05-7358-29	S	13.65	1.91	0.68
7358-148	S	29.54	1.75	1.16
Rpi05-7358-47	S	28.12	1.84	1.27
Rpi05-7358-10	S	15.26	1.58	0.70
7358-213	S	14.33	1.54	0.79
7358-276	S	22.28	1.73	1.00
Rpi05-7358-11	S	14.85	1.81	0.83
Rpi05-7358-355	S	19.07	1.68	0.87
7358-306	S	19.07	1.44	0.80
7358-S3	S	23.86	1.81	0.98

Source: Vossen, unpublished data

Appendix II NGS data of susceptible cultivars used in present study

Yield (Gbase)	Sample Name	Sample ID	Location in root file
39,560	Bzura	FR10302526	/media/bulk_01/projects/Potato_Wart/HMFreg0067_WUR-004/data/
38,302	Desiree	FR10302521	/media/bulk_01/projects/Potato_Wart/HMFreg0067_WUR-004/data/
35,766	Kuras	FR10302520	/media/bulk_01/projects/Potato_Wart/HMFreg0067_WUR-004/data/
33,342	Ludmilla	FR10302512	/media/bulk_01/projects/Potato_Wart/HMFreg0067_WUR-004/data/
37,738	VR808	FR10302518	/media/bulk_01/projects/Potato_Wart/HMFreg0067_WUR-004/data/
39,339	Bintje	Bintje	/media/scratchpad_01/essel002/Jack/Bintje
43,693	Atlantic	Atlantic	/media/scratchpad_01/essel002/Jack/Atlantik
40,685	JV18	JV18	/media/scratchpad_01/essel002/Jack/JV18
31,862	JV19	JV19	/media/scratchpad_01/essel002/Jack/JV19

Appendix III Variant filtering criteria used during variant calling

Filtering features	Criteria
Ploidy	1 or 4 (depending upon source)
Minimum mapping quality	10
Minimum base quality	10
Theta	0.01
Minimum alternate count (AO)	7
Minimum alternate fraction (AF)	0.2 or 0.12 (depending upon ploidy)
Maximum complex gap	75
Haplotype length	50
Minimum supporting mapping qsum	10
Min coverage	7
Quality (QUAL)	>40
Others (genotype qualities, use-reference allele, pooled-continuous, no-partial observation)	

Appendix IV Sequence quality for resistant bulk, susceptible bulk and resistant parent samples

Parameters for quality checking	Resistant parent	Resistant bulk	Susceptible bulk	Remarks

General statistics	Duplication = 8.4-10.5%
	GC = 34-35%
		Duplication = 6.5-8.5%
		GC = 37-38%
			Duplication = 6.1-8.1%
			GC = 37-38%	

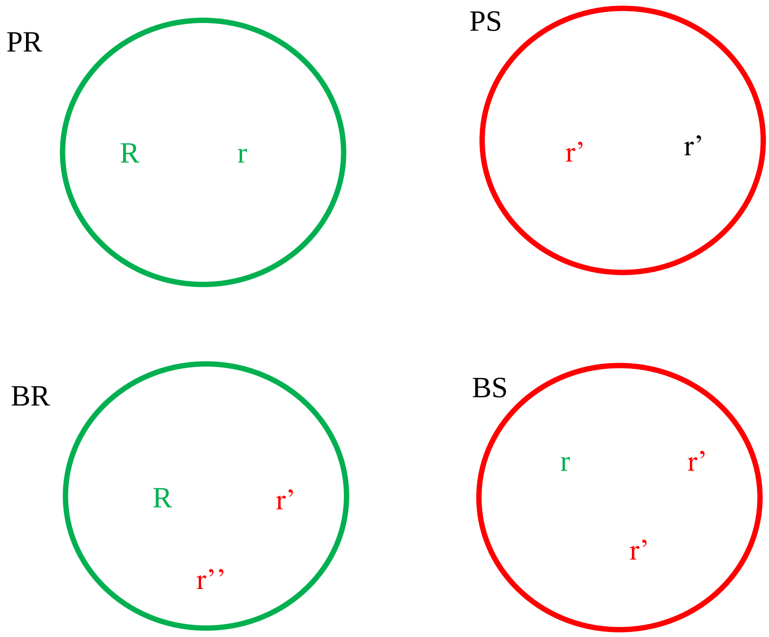
Sequence quality histograms	7/14 samples passed 7/14 samples with warnings	7/14 samples passed 7/14 samples with warnings	7/14 samples passed 7/14 samples with warnings	Phred Scores > 30 except last few base pairs in all samples
Per-sequence quality scores	14/14 samples passed	14/14 samples passed	14/14 samples passed	
Per base sequence content	14/14 samples passed	14/14 samples passed	14/14 samples passed	
Per-sequence GC content	14/14 samples passed	14/14 samples failed	14/14 samples passed	Problem in resistant bulk
Per base N content	14/14 samples passed	14/14 samples passed	14/14 samples passed	
Sequence length distribution	14/14 samples passed	14/14 samples passed	14/14 samples passed	All samples have sequences of length 151 bp
Sequence duplication levels	14/14 samples passed	14/14 samples passed	14/14 samples passed	
Over-represented sequences	14/14 samples passed	14/14 samples passed	14/14 samples passed	<1% of reads made of overrepresented sequence
Adapter content	14/14 samples passed	14/14 samples passed	14/14 samples passed	

There were 14 samples for each bulk and resistant parent

Appendix V K-mers statistics retrieved from histogram associated with resistant bulk, susceptible bulk and resistant parent samples

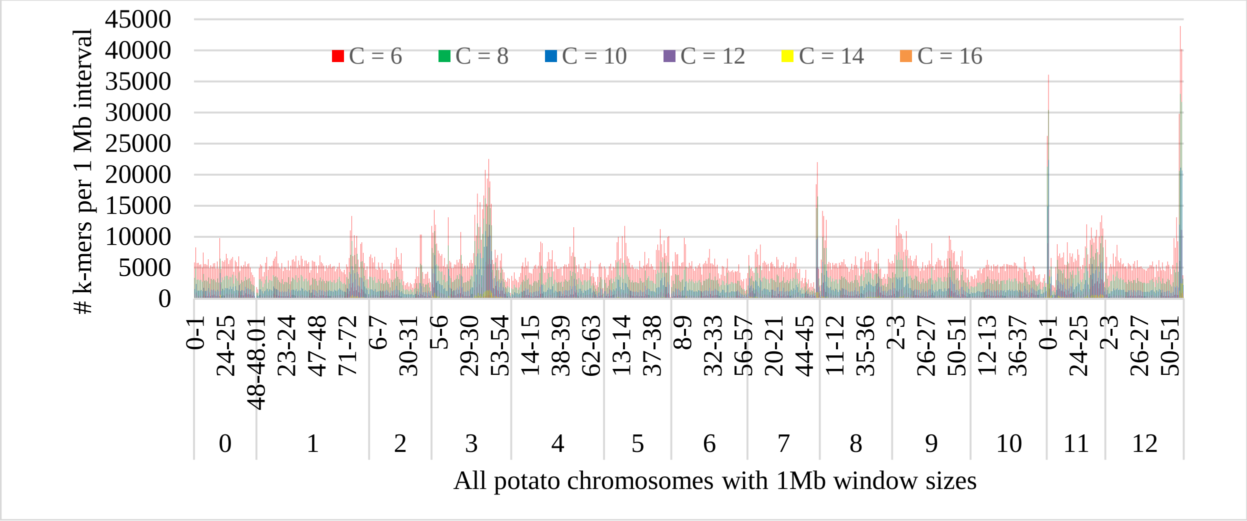
Bulks or parent	Coverage frequency					
	2-3	4-50	51-100	200-320	101-199 and >200	Total
Resistant parent						
Unique k-mers (N unique)	54,760,289	642,011,584	184,539,56	1,460,825	7,537,066	724,223,720
Total k-mers (N total)	132,611,135	13,540,499,684	1,254,372,584	363,732,964	4,334,208,973	19,625,425,340
Unique k-mers (%)	7.56	88.64	2.55	0.20	1.04	100
Total k-mers (%)	0.67	68.99	6.39	1.85	22.08	100
Resistant bulk						
Unique k-mers (N unique)	103,301,863	885,215,267	25,704,393	5,594,464	9,669,106	1,029,485,093
Total k-mers N total)	238,743,396	15,293,369,553	1,784,193,907	1,426,720,466	4,783,973,039	23,527,000,361
Unique k-mers (%)	10.03	85.98	2.49	0.54	0.93	100
Total k-mers (%)	1.01	65.00	7.58	6.06	20.33	100
Susceptible bulk						
Unique k-mers (N unique)	99,643,712	850,679,682	13,421,934	1,330,490	17,336,029	982,411,847
Total k-mers N total)	238,308,140	12,150,578,357	908,785,688	33,1977,798	3,995,098,006	17,624,747,989
Unique k-mers (%)	10.14	86.59	1.37	0.14	1.76	100
Total k-mers (%)	1.35	68.94	5.16	1.88	22.67	100

Appendix VI Expected haplotype in resistant parent, susceptible parent, resistant bulk and susceptible bulk



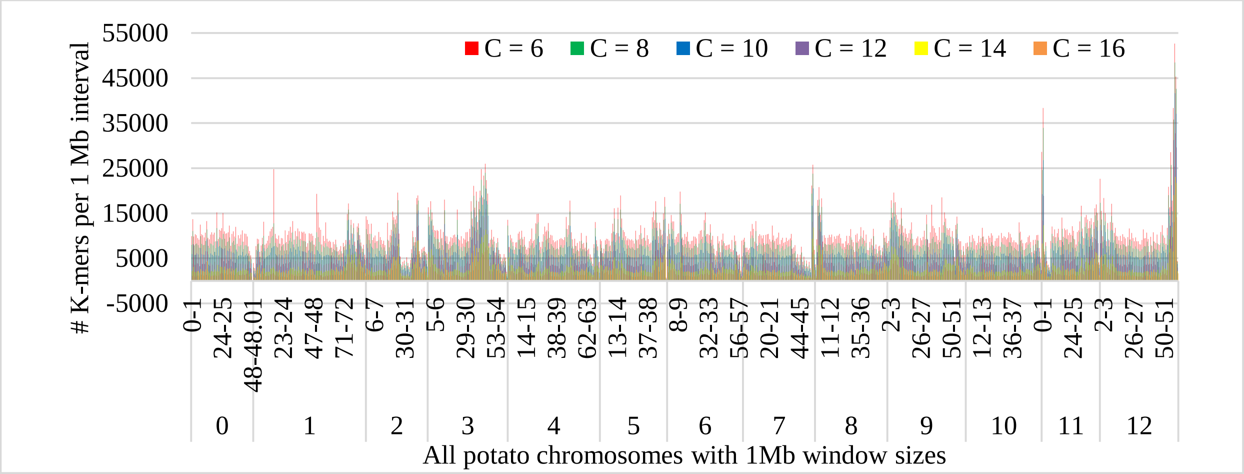
PR = resistant parent, BR = resistant bulk; BS = susceptible bulk; PS = susceptible parent; the capital alphabet inside the circle indicates the R haplotype whereas small alphabet indicates S haplotype

Appendix VII Resistant bulk specific k-mers shared with resistant parent but not present in susceptible bulk at different coverage threshold (BSA approach I)

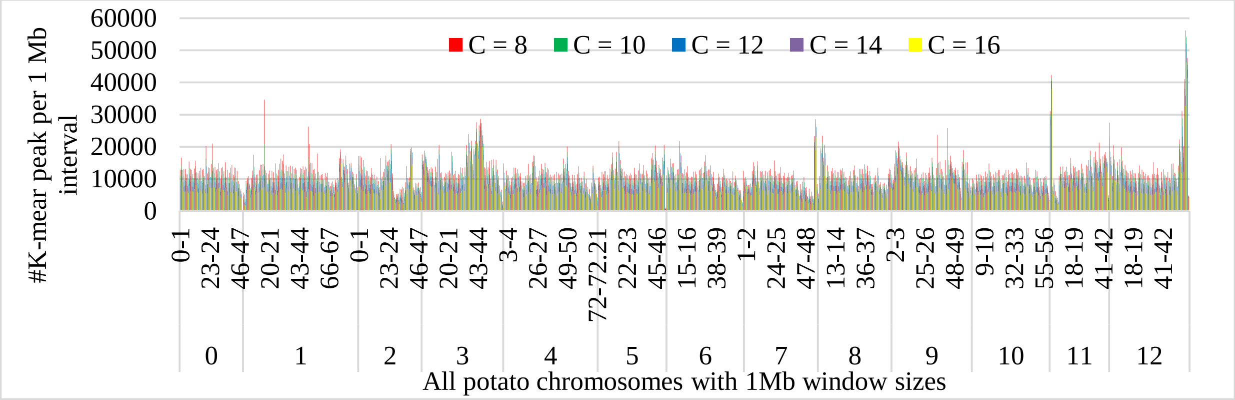


C represents the coverage threshold, for example, C 6 keep all k-mers having coverage 6 and above in final k-mers list

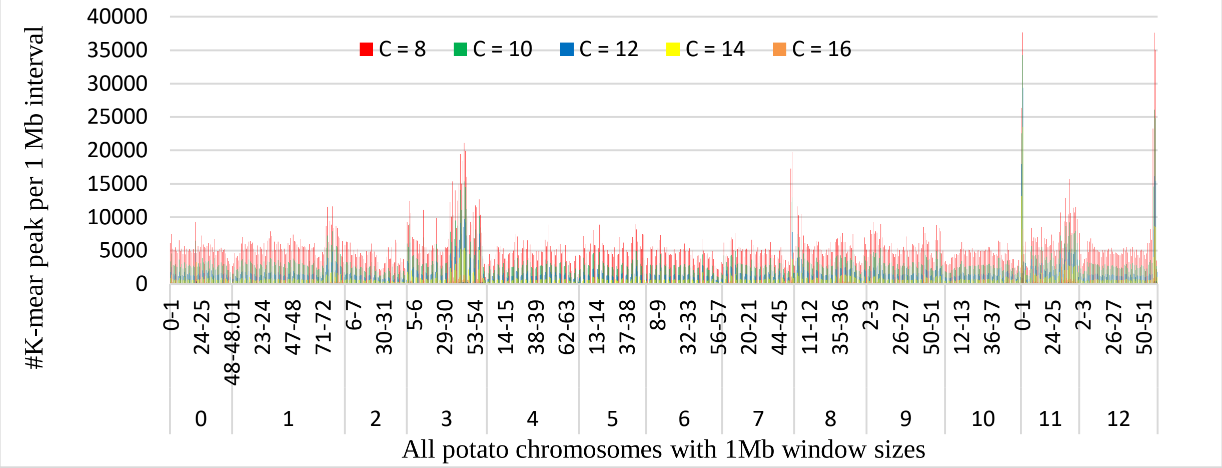
Appendix VIII K-mers in resistant parent but not present in susceptible bulk at different coverage threshold (BSA approach III)



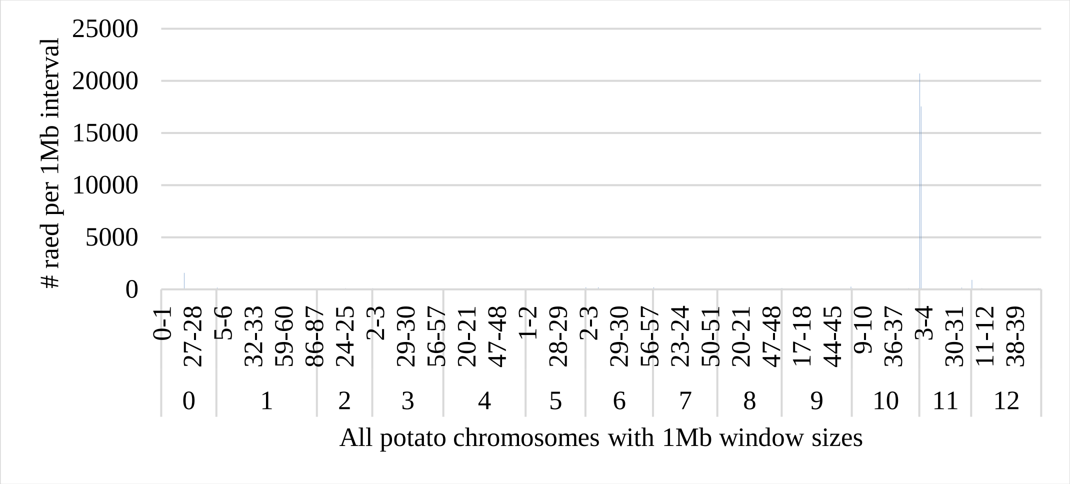
Appendix IX All k-mers in resistant parent and resistant bulk but not present in susceptible bulk at different coverage threshold (BSA approach IV)



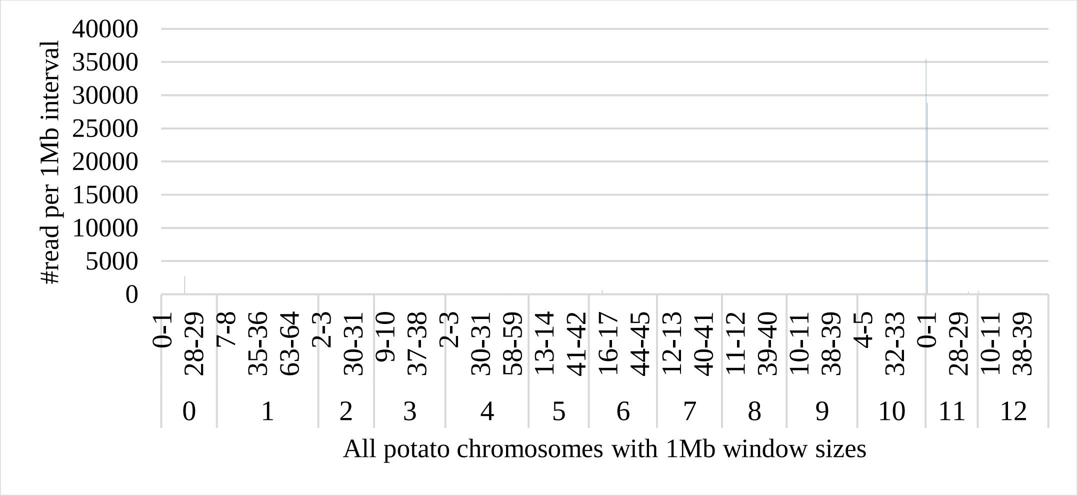
Appendix X All k-mers in resistant bulk but not present in susceptible bulk at different coverage threshold (BSA approach V)



Appendix XI Resistance haplotype-specific reads mapped to reference genome DM



Appendix XII Susceptible haplotype-specific reads mapped to reference genome DM

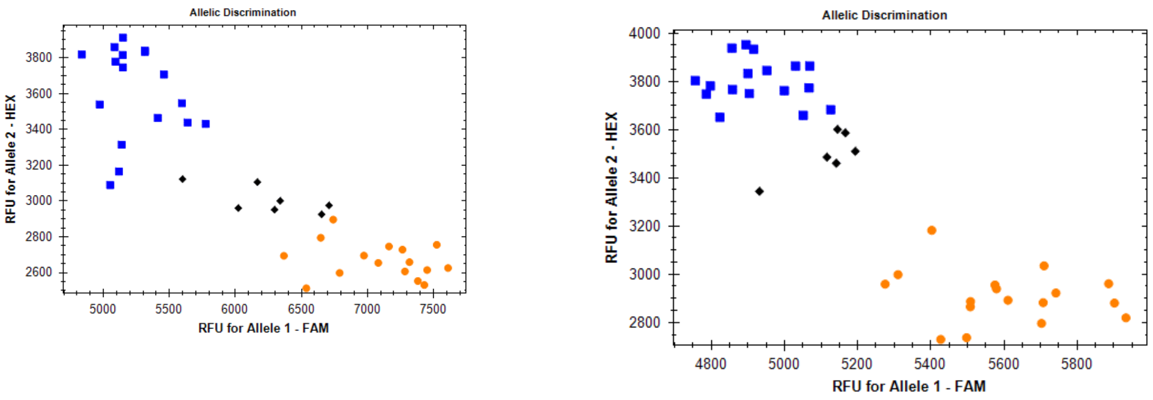


Appendix XIII Target contigs those are potential to design markers

Contigs anchored to region of interest with at least one variant						
Contigs anchored to chromosome 0	Contigs anchored to chromosome 11					
scaffold127	scaffold5	scaffold329	C877	C3269	C3804	C4074
scaffold237	scaffold8	scaffold332	C1139	C3283	C3818	C4082
scaffold306	scaffold18	scaffold335	C1339	C3339	C3858	C4090
scaffold310	scaffold21	scaffold352	C1483	C3381	C3872	C4096
scaffold366	scaffold22	scaffold370	C1561	C3407	C3878	C4112
scaffold406	scaffold107	scaffold384	C1985	C3425	C3880	C4114
C1245	scaffold128	scaffold389	C2025	C3443	C3896	C4116
C1719	scaffold145	scaffold390	C2119	C3473	C3898	C4126
C2641	scaffold158	scaffold400	C2173	C3477	C3920	C4128
C2687	scaffold167	scaffold403	C2249	C3479	C3922	C4130
C2955	scaffold182	scaffold408	C2329	C3481	C3936	C4136
C3071	scaffold184	scaffold409	C2433	C3535	C3940	C4154
C3137	scaffold192	scaffold413	C2467	C3563	C3954	C4158
C3385	scaffold199	scaffold414	C2527	C3582	C3962	C4164
C3614	scaffold203	scaffold418	C2533	C3600	C3966	C4168
C3696	scaffold207	scaffold426	C2599	C3618	C3990	C4188
C3716	scaffold212	scaffold430	C2849	C3646	C3998	C4198
C3842	scaffold226	scaffold432	C2851	C3654	C4012	C4202
C4022	scaffold234	scaffold433	C2921	C3658	C4014	C4208
C4038	scaffold243	scaffold434	C3107	C3726	C4030	C4214
C4150	scaffold252	scaffold435	C3155	C3756	C4040	C4226
C4156	scaffold258	scaffold441	C3199	C3794	C4062	C4228
	scaffold324	C733	C3249	C3796	C4072	

Region of interest represents the 0.8 to 12.5 Mb region on chromosome 11 and 20800 Kb to 20950 Kb region in chromosome 0

Appendix XIV Genotyping result using markers 9 and marker 10 on Cap7358 population and few members of Athlete x Queen Anne population

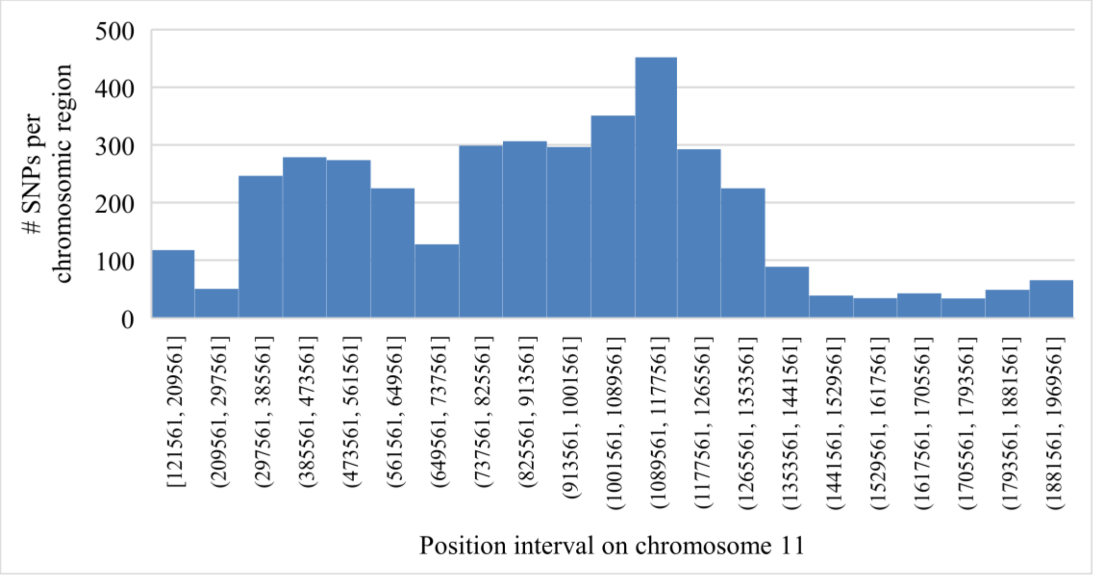


a. Genotyping result of KASP_9_7358

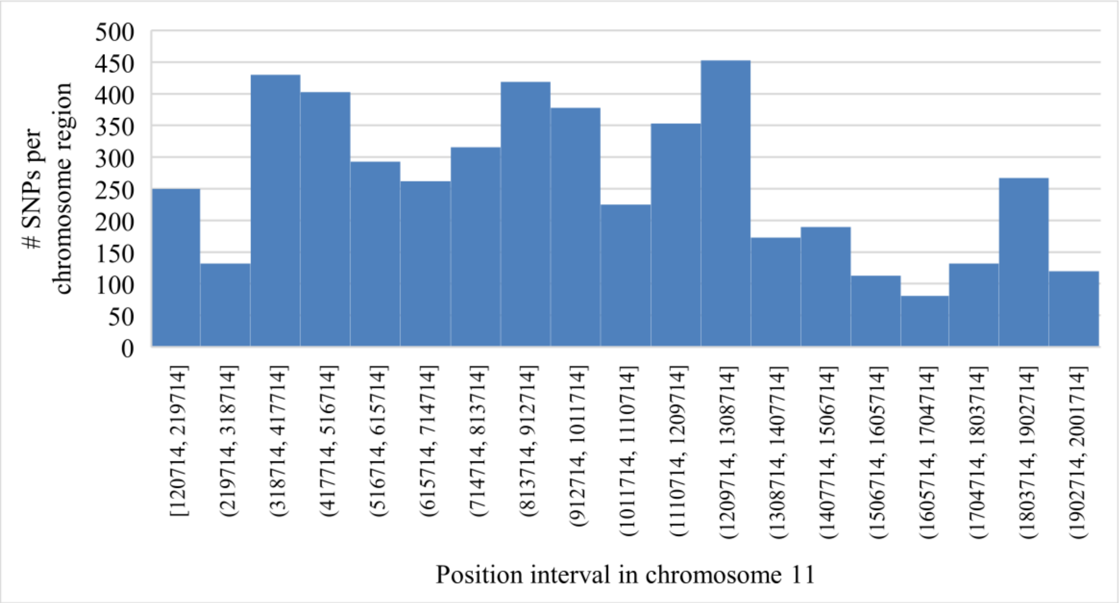
b. Genotyping maker KASP_10_7358

The resistance alleles (Allele 1) and susceptible alleles were clustered together towards x-axis and y-axis, respectively. Sample with no template DNA (MQ) and Athlete population fall in between them signifying no PCR product

Appendix XV SNPs density in resistance haplotype per chromosomic interval on chromosome 11



Appendix XVI SNPs density in susceptible haplotype per chromosomic interval on chromosome 11



If you have eaten today

Don’t forget to thank a farmers and farm workers who till the field

If you saw invisible

Don’t forget to thank scientist, who make it possible

If you felt paradigm shift on socio-economy condition of farmers

Don’t forget to thank an agriculturist who works for society, not for hours

WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018