The potato NB-LRR gene family

–

Determination, characterisation and utilisation for rapid identification of novel disease resistance genes

Florian Jupe, MSc

Doctor of Philosophy

The James Hutton Institute,
The Sainsbury Laboratory
And
University of East Anglia

December 2012

Abstract

The potato genome sequence derived from a double-monoploid *Solanum tuberosum* Group Phureja clone provides unparalleled insight into the genome composition of this important crop, including its repertoire of disease resistance genes. The vast majority of plant resistance (*R*) genes contain a nucleotide-binding and leucine-rich repeat domain, and are collectively known as NB-LRRs. The aim of this thesis is to aid the annotation of this gene family in potato and to use this information to accelerate the process of functional *R* gene isolation. The first part of this Thesis reports the identification and phylogenetic characterisation of 438 NB-LRR genes through an amino-acid motif based search of the ~39,000 potato gene-models. A full re-annotation of the single sequences provided a blueprint for a sequence capture approach to enrich for NB-LRR specific sequences prior next-generation sequencing, as reported in the second part of this thesis. In a proof-of-concept study that was carried out on the sequenced potato clone, not only 424 of the previously identified genes were verified, but further 338 NB-LRR encoding sequences were identified, mainly from un-annotated regions of the genome. Physical map positions were established for 83% of the predicted NB-LRRs across all 12 potato chromosomes.

The third part of this thesis reports on the characterisation of the genetic basis of the late blight resistance of the potato cultivar Sarpo Mira. Whole plant assays with two aggressive late blight isolates identified a differential segregation for resistance within a population derived from a cross with the susceptible cultivar Maris Piper. Transient expression of characterised late blight effector genes gave evidence for a novel resistance gene composition within Sarpo Mira. A sequencing based genotyping approach identified five loci on chromosomes 5, 11 and 12 with positive impact on the resistance. Analysis of NB-LRR enriched Illumina reads of bulked resistant and susceptible plants was however inconclusive, due to difficulties arising from the tetraploid genetic background of Sarpo Mira.

Table of Contents

List of Figures

List of Tables

Accompanying material

Annex 1 Gene bank (NCBI) accession numbers for proteins used in the positive and negative training sets. 'Positive' NB-LRR and 'negative' non-NB-LRR sequence training sets were used with the MEME Suite psp-gen script (version 4.4.0) (Bailey et al. 2010) to identify discriminative motifs from the positive set.

Annex 2 Protocol for NB-LRR capture enrichment experiment. This protocol is a composition of the manufacturer's protocols (Agilent Technologies and Agencourt), as well as the method described by Quail et al. (2009).

Annex 3 Malcolmson's 1-9 scale of increasing resistance scores. This scheme was used to score the segregating Sarpo Mira × Maris Piper population. Score 9, representing 0% necrotic tissue is not shown in the original figure from Cruickshank et al. (1982)

Digital Accompanying material

E1 List of identified DM NB-LRR genes. Identified NB-LRR genes are listed, together with information on their PGSC identity, coding DNA strand, annotation, number of identified open reading frames (ORFs), the predicted pseudomolecule (LG), start of original DMG on LG, end of original DMG on LG, repeat density, gene density, and motif complement of the annotated sequence DMG+.

E2 Detailed phylogenetic analyses of the DM NB-LRR NB-ARC domains. The NB-ARC domains of TNL and CNL type gene products were used, alongside selected NB-ARC domains from functional resistance genes, to study the phylogenetic relationships between them.

E3 List of the novel identified NB-LRR genes. This table contains all identified potato NB-LRR genes sorted according to the chromosomal position, based on the pseudomolecules version 3.1.10. This table further provides information on the coding direction of the sequences, and the NB-LRR type of the corresponding gene. The last two columns show the computed positions on pseudomolecules version 3.1.11.

E4 DM NB-LRR_cds.fasta

E5 DM NB-LRR_gene_products.fasta

E6 new NB-LRR DNA.fasta

FASTA sequences are presented for (E4) the re-annotated DM NB-LRR coding sequences, (E5) the conceptual translations, and (E6) the new identified NB-LRR encoding genomic regions from chapter 3. IDs in E4 and E5 correspond to the original DMG identifiers provided by the PGSC. IDs in E6 correspond to the respective chromosome as well as to the number of novel NB-LRR on the chromosome.

E7 Detailed views of potato chromosomes 1 - 12. Genes encoded by the positive DNA strand are depicted on the left hand side of the chromosome, whereas those encoded by the negative strand are shown on the right. NB-LRR genes belonging to clusters are indicated by vertical bars. Heterogeneous clusters are indicated by an *.

List of Abbreviations

AFLP - amplified fragment length polymorphism
ARC - human apoptotic protease-activating factor-1 (APAF-1), plant R proteins and *Caenorhabditis elegans* death-4 protein (CED-4)
Avr - avirulence gene
BAC - bacterial artificial chromosome
BAM - binary version of sequence alignment/map file
blast - basic local alignment and search tool
BR - bulked resistant plants
BS - bulked susceptible plants
BSA - bulk segregant analysis
C1 - NB-LRR cluster 1
CC - coiled coil
cDNA - complementary DNA
cds - coding sequence
cM - centi Morgan
CNL - CC-NB-LRR
CPC - Commonwealth Potato Collection
DM - doubled monoploid *S. tuberosum* Group Phureja clone DM1-3 516 R44
DMB - DM assembled superscaffold
DMG - DM gene model
DMP - DM protein model
dpi - days post inoculation
EST - expressed sequence tag
ETI - effector triggered immunity
ETS - effector triggered susceptibility
gDNA - genomic DNA
gff - general feature format
HNL - Hydrolase-NB-LRR
hpi - hours post inoculation
HR - hypersensitive response
LRR - leucine rich repeat
MAMP - microbial associated molecular pattern
MAST - motif alignment and search tool
miRNA - micro RNA
MLG - multi-locus genotype
MP - Maris Piper
NB/NBS - nucleotide binding site
NB-LRR - nucleotide binding site leucine rich repeat
NGS - next generation sequencing
PAMP - pathogen associated molecular pattern
PCD - programmed cell death
PE - paired end
PNL - protein kinase-NB-LRR
POPA - Potato Oligo Pooled Assay
PRR - pathogen recognition receptor
psp-gen - position specific priors script (-gen is unspecified)
PTI - pathogen triggered immunity
qPCR - quantitative polymerase chain reaction
QTL - quantitative trait loci
R - resistance (gene/protein)
RAD-seq - restriction site associated DNA sequencing

RFLP - restriction fragment length polymorphism
RGH - resistance gene homologue
RH - heterozygous diploid potato clone
RLK - receptor-like kinase
RLP - receptor-like protein
Rpi - resistance to *Phytophthora infestans*
SAM - sequence alignment/map
SGS - second generation sequencing
siRNA - short interfering RNA
SM - Sarpo Mira
SNP - single nucleotide polymorphism
STAND - signal transduction ATPase with numerous domains
TIR - Toll and Interleukin receptor
TNL - TIR-NB-LRR
UTR - untranslated region

Publications arising from this Thesis

Jupe, **F.**, L. Pritchard, G. J. Etherington, K. Mackenzie, P. J. Cock, F. Wright, S. K. Sharma et al. 2012. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC.Genomics* 13:75.

Jupe, **F.**, K. Witek, W. Verweij et al. RenSeq: sequencing DNA enriched for plant resistance genes enables re-annotation of sequenced genomes and rapid mapping of resistance loci in segregating populations. In preparation.

Jupe, **F.**, C. Hackett et al. Investigations into the late blight resistance of potato cultivar Sarpo Mira. In preparation.

Acknowledgements

During three tremendously quickly passing years of my PhD study, I met many people of which some were involved in work presented in this thesis, and I am very grateful to all of them.

Most important, I want to say thanks to my supervisors Dr Ingo Hein, Prof Jonathan Jones and Dr Glenn Bryan for arranging this project and supporting me throughout.

A big thanks to Philip Smith for proof-reading this thesis and other manuscripts in the past.

I have to thank the James Hutton Institute and The Sainsbury Laboratory for the financial support of this joint PhD-studentship.

I want to thank all past and present members in Ingo's lab, especially Brian for his support with blight experiments, Pauline, Gaetan, Zul and Chen for all the fun in the lab. Furthermore a big thanks to all members of Jonathan's lab, but mainly Walter and Kamil for their company on the potato project.

Very special thanks go to Stefan and the table football for all the hours that we spent together playing, discussing and gossiping.

Scotland, thank you for your relaxing and mind refocusing beauty.

Thank you, my dear Julietta, for joining me on this journey, and I'm looking forward to all the journeys ahead of us!

Ein grosser Dank an meine und Julie's Familie fuer die Unterstuetzung waehrend all den Jahren des Studiums. Nia grod'aus, imma was eigens, aba trotzdem imma wos Bsonders.

“Although the subject material of this research was my own choice, at the time it was determined on I was quite ignorant of the very special advantages as well as disadvantages which the Potato offers for the Mendelian student.”

- Redcliffe N. Salaman (1911) -

Chapter 1 – General Introduction

1.1 The plant innate immune system

Plants have evolved a sophisticated, multi-layered defence network to detect and respond to pathogen challenges (Jupe et al. 2012). The first non-inducible layer is composed of physical barriers, such as the waxy cuticles, which are sufficient to hold off most pests and pathogens (Freeman and Beattie 2008). In addition, non-inducible responses can include antimicrobial compounds that are pre-formed prior to pathogen challenge and are thus referred to as phytoanticipins (VanEtten et al. 1994). Pathogens that are successful in penetrating this first layer, are then confronted by two layers of inducible defences according to current knowledge of the plant innate immune system (Jones and Dangl 2006; Shivaprasad et al. 2012;Tameling and Takken 2008). Jones and Dangl (2006) illustrated the inducible plant immune system quantitatively against the amplitude of defence in the "Zig-Zag-Zig"-model over three phases, which arguably represent the levels of co-evolution. This model contains also a fourth phase, shown for the interaction of *Solanum* species with the oomycete pathogen *Phytophthora infestans* (Hein et al. 2009b) (Figure 1). The resulting model is however generic and applicable to other (hemi-) biotrophic pathogens.

Phase 1: PAMPs and Elicitors

In the first layer of the inducible plant immune system, specialised pattern recognition receptors (PRRs) at the plant cell surface sense microbial- or pathogen associated molecular patterns (MAMPs or PAMPs). This recognition confers pathogen triggered immunity (PTI) through downstream signalling processes leading to a defence reaction (Boller and He 2009;Chisholm et al. 2006;Cui et al. 2009;Jones and Dangl 2006). MAMPs or PAMPs are microbial components that are of fundamental biological importance for the pathogen and widely distributed among them, however absent in the host. Among them are bacterial flagellin, elongation factor TU, lipopolysaccharides, heptaglucosides, or chitin from fungal or insect exoskeletons (Boller and Felix 2009;Hein et al. 2009b;Jones and Dangl 2006).

A major group of PRRs are the receptor like kinases (RLKs) which consist of an extracellular leucine rich repeat (LRR)-domain, a transmembrane domain and an additional kinase domain into the cytoplasm (Dangl and Jones 2001). The first evidence for PRR-mediated PTI was shown for the recognition of a 15-22 amino acid long conserved domain of the bacterial flagellin, which was later found to be recognised by

the RLK *Flagellin Sensing 2* (FLS2) in *Arabidopsis* (Felix et al. 1999). Mutants lacking FLS2 showed an increased susceptibility towards virulent bacteria (Chinchilla et al. 2006;Zipfel et al. 2004).

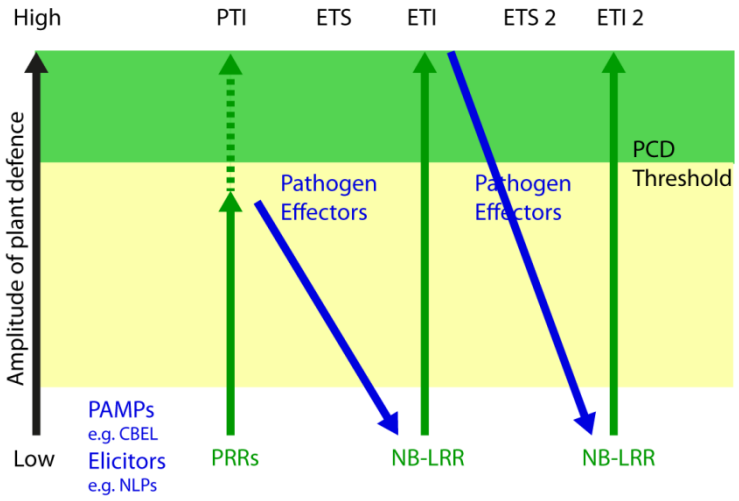


Figure 1 The Zig-Zag-Zig in oomycete-plant interactions, modified from Hein et al. (2009b) and Jones and Dangl (2006). The model shows the four phases of inducible plant immunity as a representation of co-evolution, over the amplitude of plant defences. Oomycete specific pathogen associated molecular patterns (PAMPS) and elicitors are shown, that can trigger pathogen recognition receptors (PRRs) mediate PAMP-triggered immunity (PTI). Successful pathogens have evolved effector molecules that can suppress this plant defence leading to effector triggered susceptibility (ETS), but can also lead to recognition by plant resistance (R) proteins resulting in effector triggered immunity (ETI).

Phase 2: pathogen effectors modulate the plant immune system

While PTI stops the unsuccessful pathogens from infecting plants, adapted pathogens are able to suppress this response by interfering with downstream signalling components of the resistance pathways. So called 'effector' molecules are released into the intracellular space suppressing extracellular defence-related proteins, or are actively introduced into the host-cells to interfere with downstream signalling of plant defence responses and therefore to promote virulence. These proteins are thus key features of the pathogen and might have evolved during the co-evolution with its hosts (Hein et al. 2009b;Kelley et al. 2010). Examples for different effector functions are the *P. infestans* effector Avr3a, that mediates intracellular suppression of INF1 induced cell death (Bos et al. 2010;Schornack et al. 2009), or the extracellular *Cladosporium fulvum* protein Avr4 which was shown to inhibit plant chitinases and thus prevents hydrolysis of the fungal cell wall (van den Burg et al. 2006). A further *C. fulvum* example is its LysM effector Ecp6

that prevents PTI through the sequestration of chitin oligosaccharides that detach from its own cell walls (de Jonge et al. 2010).

Phase 3: Recognition of the effector

The second layer of induced defences is based on successful recognition of these effectors or changes of their targets, by the cognate resistance (R) protein and results in effector-triggered immunity (ETI). Jones and Dangl (2006) described the ETI as an accelerated PTI that often results in a hypersensitive response (HR), a form of a programmed cell death at the site of pathogen detection. The most numerous group of *R* genes in plant-microbe interactions comprises proteins sharing a nucleotide binding-site followed by a LRR-domain (NB-LRR). This group will be discussed in more detail in inter section *1.1.2*. Another ETI-conferring protein family is the receptor-like proteins (RLP), which are structurally very similar to RLKs. Prominent examples are the tomato genes *Cf-4* and *Cf-9*, recognising the *Cladosporium fulvum* effectors *Avr4* and *Avr9*, respectively, or *Ve1* recognising the *Verticillium Ave1* gene (de Jonge et al. 2012;Hammond-Kosack et al. 1994;Joosten et al. 1994;Kawchuk et al. 2001). These receptor-like proteins possess a transmembrane domain at their C-termini and an extracellular LRR; also they lack the kinase domain common to RLKs (Rivas and Thomas 2005).

Another small group of R proteins consists of cytoplasmic Serine/Threonine kinases. An example is Pto which is able to form a complex with the NB-LRR Prf to recognise the effectors AvrPto and AvrPtoB. Similar complexes have been shown for other non-Pto kinases (Dangl and Jones 2001;Gutierrez et al. 2010).

Phase 4: Natural selection and co-evolution

In phase 4, driven by natural selection, pathogens have evolved various means to interfere with ETI. Examples include the diversification of recognized effectors, the delivery of suppressors of programmed cell death (PCD) or utilising functional redundancy within the effector repertoire, discarding recognised effectors or blocking their expression (Hein et al. 2009b;Wang et al. 2011).

For the *P. infestans* effector Avr3a, Armstrong and colleagues suggested a positive selection through gene duplication to avoid allelic recognition, and a similar diversifying selection has also been reported for the cognate *R3a* gene (Armstrong et al. 2005;Huang et al. 2005). Another recent finding suggests that the nematode effector SPRYSEC-19 has evolved to suppress PCD after ETI through interaction with the LRR domain of a member

of the *Sw-5* NB-LRR gene family, known to confer viral resistance (Brommonschenkel et al. 2000;Postma et al. 2012).

1.1.1 Models for *R* gene function

During the first half of the 20th century, experiments on incompatible plant–microbe interactions led Harold Flor to conclude a "gene–for–gene" interaction, in which a pathogen derived gene product is sensed by a plant protein leading ultimately to a resistance response (Flor 1942). Research over the years proved this model to be correct in several cases, but expanded it further to include "genes–for–gene", "gene–for–genes", and "genes-for-genes" interactions (reviewed in Gassmann and Bhattacharjee (2012)). A further form of interaction is postulated in the so called guard–hypothesis, in which R proteins monitor the status of effector targets, rather than the effector itself (Gassmann and Bhattacharjee 2012;Jones and Dangl 2006;Thomma et al. 2011;Van der Biezen and Jones 1998a).

An example for a direct gene–for–gene interaction was shown by Krasileva et al. (2010), who reported that the LRR domain of the *Arabidopsis* NB-LRR protein RPP1 mediates interaction with its cognate effector ATR1. A recent example is the direct recognition of the effectors IPI-O1 and IPI-O4 through the coiled coil-domain of the potato *R* gene *RB/Rpi-blb1* (Chen et al. 2012). The *Nicotiana tabacum* NB-LRR *N* was shown to only function properly in conjunction with the NB-LRR *NRG1* (*N required gene 1*), and may thus be an example of a genes–for–gene interaction (Peart et al. 2005). However it may also be that only one NB-LRR is responsible for the recognition, while the other functions in signalling (then gene-for-gene interaction).

In many cases, *R* genes confer resistance after elicitation by specific *Avr* genes, however a direct interaction was not found. These results suggest an indirect interaction, and gave a first indication for the guard hypothesis (Dangl and Jones 2001;Gassmann and Bhattacharjee 2012;Jones and Dangl 2006;Takken and Tameling 2009;Tameling and Takken 2008;Van der Biezen and Jones 1998a). The *Arabidopsis* protein RIN4 is targeted and phosphorylated by a number of diverse *Pseudomonas syringae* effectors like AvrRpm1 and AvrB, and these phosphorylations in turn activate the NB-LRR RPM1 (Deslandes and Rivas 2012;Liu et al. 2011;Mackey et al. 2002). A further example was reported by DeYoung et al. (2012), whose experiments showed that the *Pseudomonas syringae* effector AvrPphB targets and subsequently cleaves the *Arabidopsis* protein PBS1, and that these cleavage products in turn activate the NB-LRR RPS5 which elicits ETI.

1.1.2 *R* genes – The NB-LRR protein family

As mentioned earlier, the most numerous class of *R* genes contains proteins with a nucleotide-binding site (NB) and leucine-rich repeat (LRR) domain, and are members of the STAND (Signal Transduction ATPase with Numerous Domains) protein family of NTPases, known as NB-LRRs (Lukasik and Takken 2009;Van der Biezen and Jones 1998a). The nucleotide binding site forms part of a larger complex known as NB-ARC, which reflects its presence in the human apoptotic protease-activating factor-1 (APAF-1), plant R proteins and *Caenorhabditis elegans* death-4 protein (CED-4) (Van der Biezen and Jones 1998b). Further subdomains and multiple conserved motifs have been identified within the NB-ARC domain (Lukasik and Takken 2009). The LRR domain is involved in the protein/protein interaction (Padmanabhan et al. 2009;Takken and Tameling 2009;Tameling and Takken 2008), and might also act as a repressor to prevent inappropriate NB activation (Belkhadir et al. 2004). A detailed review on the structurally diverse N-termini of NB-LRRs can be found in Tameling and Takken (2008). Based on the presence or absence of N-terminal domains, members of the NB-LRR family can be divided into two major groups. The first group contains an N-terminal domain with homology to the *Drosophila* Toll and human interleukin-1 receptor (TIR); this group is referred to as TIR-NB-LRRs or TNLs (Figure 2). The second, non-TIR-NB-LRR, group is collectively known as CNLs as some, but not all, members of this group contain a predicted coiled-coil (CC) structure in the N-terminus (Figure 2). A further, rather small group with three members in *Arabidopsis* also contains a WRKY transcription factor binding site (Noutoshi et al. 2005). Recently, Xue et al. (2012) described two further groups of NB-LRRs in Bryophytes, PK-NB-LRR (PNL) harbouring an N-terminal protein kinase domain, identified from the moss *Physcomitrella patens,* and Hydrolase-NB-LRR (HNL) from the liverwort *Marchantia polymorpha* (Figure 2). The division of NB-LRR proteins is, however, not only manifested in the N-terminal structure, but also in the sequence of the NB-ARC domains as is shown in various phylogenetic analyses (Jupe et al. 2012;McHale et al. 2006;Meyers et al. 1999;Meyers et al. 2002;Xue et al. 2012).

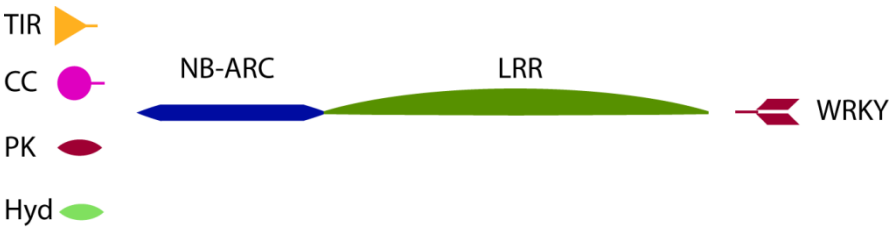


Figure 2 NB-LRR proteins share a conserved NB-ARC domain followed by leucine rich repeats in the C-terminus. The most common N-terminal domains are TIR (Drosophila Toll and human like Interleukin-1 receptor) and CC (coiled coil). Only recently the N-terminal domains PK (protein kinase) and Hyd (hydrolase) were identified from two moss species. In Arabidopsis, a low number of TIR-NB-LRR is found with an C-terminal WRKY binding domain.

As stated in Jupe et al. (2012), NB-LRR genes comprise one of the most numerous gene families in plants, and recent plant genome sequencing projects allowed the identification of the NB-LRR gene complements in an array of plant species. Approximately 150 NB-LRR encoding genes have been identified in the genome of *Arabidopsis thaliana* Col-0 (Meyers et al. 2003), 185 within *Arabidopsis lyrata* (Guo et al. 2011), 92 within *Brassica rapa* (Mun et al. 2009), 416 and 535 in the genomes of the woody species poplar and grapevine respectively (Yang et al. 2008b), and 464 and 483 in two genomes of rice *Oryza sativa* (Yang et al. 2008a). In addition, partial NB-LRRs that lack some NB-LRR specific domains and contain, for example, only TIR, TIR-NB, CC, and CC-NB domains, have been described in plant genomes (Guo et al. 2011;Meyers et al. 2002). NB-LRR genes are ancient in their origin and have been identified in ancestors of early land plants, as a recent analysis of bryophyte genomes has shown (Xue et al. 2012). TNLs and CNLs have been found in the genomes of a wide range of gymnosperms and eudicots (Tarr and Alexander 2009). However, the composition of NB-LRR genes varies significantly between species (Cannon et al. 2002). The unequal representation of NB-LRR lineages within plant taxa has been typified by the low frequency of TNLs within monocotyledonous species despite the manifestation of TNLs prior to the angiosperm–gymnosperm split (Jiang et al. 2005;Tarr and Alexander 2009; as stated in Jupe et al. 2012).

NB-LRR genes are organized either as isolated genes or as linked clusters of varying size that are thought to facilitate rapid *R* gene evolution (Hulbert et al. 2001). NB-LRR gene clusters are termed homogeneous when they contain only sequences that share a recent common ancestor. In contrast, clusters that contain more distantly-related NB-LRRs are referred to as heterogeneous (Friedman and Baker 2007).

1.1.3 NB-LRR gene expression is tightly regulated

The ability to induce cell death leads to a necessity for a tight regulation in the plant cell (Huang et al. 2005;Shen et al. 2002;Tan et al. 2008). The auto inhibition of NB-LRR genes is mainly provided by intramolecular interactions between the different domains (Takken and Tameling 2009), or as recently discovered through micro (mi)RNA-based post-transcriptional gene silencing (Li et al. 2012;Shivaprasad et al. 2012). Two independent studies highlight miRNA families that are present in Solanaceae plant species and specifically target the P-loop motif of NB-LRR genes, initiating a gene silencing cascade of NB-LRRs through secondary short interfering (si)RNAs (Li et al. 2012;Shivaprasad et al. 2012). In turn, infection of plants with viruses or bacteria results in a rapid decrease of miRNA levels and thus initiates an increase in *R* gene transcript accumulation. This decrease in miRNA levels is hypothesized to be triggered by pathogen silencing suppressors (Li et al. 2012;Shivaprasad et al. 2012). Thus, the regulation of *R* genes via miRNAs provides an independent means to detect pathogen effector activity and potentially bring about resistance.

1.2 The tuber crop potato

Potato is with a worldwide production of 324 million tonnes in 2010 the most produced non-cereal staple food crop (http://faostat.fao.org/site/339/). Now a global crop, the potato was only introduced to Asia, Africa, Europe and North America when it was discovered by Spanish conquistadores during the 16th century in South America. Herbalists like Carol Clusius spread the potato Europe-wide, initially as a botanical specimen rather than a food source (Hawkes 1993). Introductions of potatoes into Africa, China, India, and Japan took place during the 17th century by British missionaries. The first plant in Europe was the short-day adapted *S. tuberosum* Group Andigena, originating from the Andes. Only able to form tubers during the winter months in the frost-free zones of Spain and Italy, it took several centuries of selection for tuberisation under long-days to spread into northern and eastern Europe as a food source allowing large scale field production during the early 1800s (Hawkes 1993).

The cultivated Potato and its wild relatives are members of the very diverse family Solanaceae, that further hosts, amongst others, tomato (*Solanum lycopersicum*), tobacco (*Nicotiana tabacum*), and pepper (*Capsicum annuum*). The main potato species grown in the northern hemisphere is the autotetraploid *S. tuberosum* Group Tuberosum, which is a tetraploid descendant of the tetraploid *S. tuberosum* Group Andigena, member of the section Petota. The origin of this was suggested to be a result of

chromosome doubling of the diploid *S. stenotomum*, or a hybridisation with the diploid *S. sparsipilum* (Hawkes 1993), while the hybrid theory was supported by more recent DNA sequence analyses (Rodriguez et al. 2010). The ploidy levels of wild potato relatives range between diploid (2n=2x=24) and hexaploid (2n=6x=72).

1.2.1 Important Potato germplasm collections

Within the section Petota, single species have evolved in a wide range of natural habitats (Hawkes 1993;Rodriguez et al. 2010;Spooner and Clausen 1993). These various sources of natural diversity are an invaluable source for various traits including biotic and abiotic stress resistance, or tuber development that can be exploited in breeding programmes. Therefore a wide range of wild and cultivated potato species are stored in genebanks around the world. Prominent examples among many are the Commonwealth Potato Collection (CPC) in the UK, the International Potato Center (CIP) in Peru, and the Pavlovsk Research Station in Russia, but also the Dutch-German Potato Collection (DGN, Wageningen, The Netherlands).

Initiation of the CPC took place after identification of the late blight resistance of the hybrid *S.* × *edinense,* a hybrid between cultivated potato and the wild relative *S. demissum*. The CPC maintains approximately 1500 accessions from 80 wild and cultivated potato species, most collected during *The British Empire Expedition* in 1939 (Hawkes 1993) (http://germinate.scri.ac.uk/germinate_cpc/app/index.pl). Nowadays, the CPC is part of a network of international potato genebanks and is situated at the James Hutton Institute in Dundee, Scotland.

1.2.2 The potato genome sequence

In 2011, the draft genome sequence of the doubled monoploid *S. tuberosum* Group Phureja clone DM1-3 516 R44 (hereafter referred to as DM) was published by The Potato Genome Sequencing Consortium (PGSC et al. 2011). The assembly comprises 844Mb sequence information from which approximately 39,000 genes were predicted. A total of 86% of the assembly could be anchored to 12 chromosomes to create pseudomolecules. The remainder is situated on yet unanchored superscaffolds. Using a doubled monoploid clone prevented difficulties arriving from high levels of heterozygosity that are associated with ploidy levels of all types of potato (PGSC 2011). Comparative analysis with a heterozygous diploid clone showed that inbreeding

depression, shown in the reduced vigour of DM, might be caused by a number of deleterious mutations such as premature stop codons that are distributed widely over the genome. From the gene annotations 408 NB-LRR like sequences were identified. RNAseq analysis was used in this study to identify genes that are specifically regulated during tuber development, a trait that is exclusive to the section Petota (PGSC 2011). In addition to the potato genome, assemblies of the domesticated tomato *Solanum lycopersicum* and its wild relative *S. pimpinellifolium* were released and compared (TGC 2012). Within the family Solanaceae a 100 genome project has been initiated with the aim to create physical and genetic maps to further investigate the origin of culturally important traits (http://solgenomics.net/organism/sol100/view).

1.3 The late blight pathogen *Phytophthora infestans*

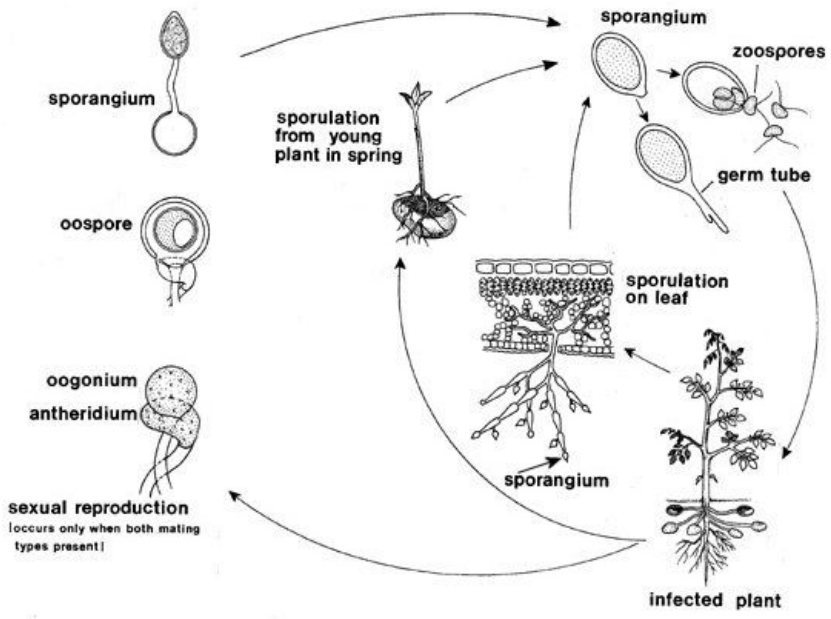
Phytophthora infestans is the causal agent for the devastating potato and tomato late blight disease and was responsible for the infamous Irish potato famine in the 1840s. The disease is caused by an oomycete pathogen and remains even more than 160 years after the famine the most destructive disease of potato and tomato world-wide, causing conservatively calculated crop losses of 6.7 billion US$ per year (Fry 2008;Haas et al. 2009;Haverkort et al. 2008).

1.3.1 The life cycle of *P. infestans*

Phytophthora is a hemi-biotrophic eukaryotic pathogen from the kingdom Chromista and the family Oomycetes. Although it shows fungus like growth structures, oomycetes are phylogenetically and physiologically very distinct organisms (Rossman and Palm 2006). Not only have DNA sequence analyses shown that oomycetes are more closely related to brown algae, but analyses show that the outer cell wall is composed of beta glucans and cellulose, compared to chitin in fungi (Rossman and Palm 2006). The ability for an asexual life-cycle is responsible for an explosive growth after primary infection of plant tissue with air dispersed sporangia, which contain on average eight flagellated zoospores (Fry 2008;Shattock 2002)(Figure 3). Infection of leaf tissue begins with germination of sporangia or zoospores and formation of an appressorium from where the leaf cuticle is penetrated by a hyphae-like structure. From these intercellular growing structures, haustoria are pushed into mesophyll cells without destroying the host cell membrane (Greenville-Briggs and Van West 2005). These haustoria are thought to mediate the exchange of nutrients as well as immuno-modulating factors such as

effectors during the biotrophic life stages of *P. infestans*. After five to seven days this close interaction comes to a local halt resulting in necrotisation of the infested tissue and the appearance of sporangiophores on the abaxial side of the leaf while growth continues on the leading infection site (Greenville-Briggs and Van West 2005). This rapid reproduction can cause a complete crop loss within seven to ten days (Fry 2008). Sporangia can be transmitted on to tubers and infect these during the storage or the emerging shoots in the next growing season.

P. infestans is also able to reproduce sexually through interacting mycelia of two different mating types (Figure 3). The emerging oospores were shown to be able to survive up to 4 years in the ground (Turkensteen et al. 2000). In contrast to South America, the European population is however dominated by clonal lineages derived from asexual reproduction despite the presence of both mating types (Cooke et al. 2012).



This is a simplified disease cycle for late blight of potato.

Figure 3 The life cycle of *Phytophthora infestans* begins with sporangia that germinate either directly or release approximately eight flagellated zoospores which are able to penetrate the intercellular space of plant tissue. During this time the host is kept alive as a food source, but dies during the development of sporangia in the asexual life cycle. The sexual life cycle occurs when both mating types are present and results in the formation of oospores (reproduced from Schumann and D'Arcy 2000).

1.3.2 Control of *P. infestans*

As mentioned before, *P. infestans* is the most destructive pathogen of potato and tomato, and control of late blight epidemics is very important for growers of both crops world-wide. Control measures taken by farmers include the application of fungicides, adaptation of cultural practices and the deployment of resistant plant material. The use of fungicides can require weekly applications during weather conditions that cause high late blight pressure. However, their use has recently been restricted by the ban of certain compounds by the European Union. Furthermore, isolates have emerged that are resistant to certain pesticide compounds.

Periods of high risk for blight infections are known as Smith Periods (Smith 1956), and consist of at least two consecutive days where the minimum temperature is 10°C or above and the relative humidity on each day is greater than 90% for at least 11 hours (from www.Blightwatch.co.uk). Free notifications of these periods are being offered by Blightwatch via email or SMS. From an economic and ecological point of view however, the most preferable solution is based on a genetic and physiological resistance of the cultivated plants towards the late blight pathogen (Goverse and Struik 2009).

1.3.3 Dynamics within the population of *P. infestans*

Until about 1976, the European population of *P. infestans* consisted of only the A1 mating type (Fry, 2008). The first discovery of an A2 mating type was made in 1984 in Switzerland. There is evidence that the A2 mating type arrived with a contaminated potato shipment from Mexico in 1976 (Goodwin and Drenth 1997). Since the middle of this decade, a change in the western European *P. infestans* population has been noted. A new A2 type of *P. infestans*, the multi-locus genotype (MLG) 13_A2, has become the dominating genotype between 2006 and 2010, accounting for up to 79% of all UK field isolates in 2008 (Cooke et al., 2012)(Figure 4). Isolates of this genotype show high levels of aggressiveness and are able to overcome the resistance of many potato varieties including 'Stirling' and 'Lady Balfour' (Fry, 2008; Cooke et al. 2012). Experiments with the MLG 13_A2 isolate 2006_3928A showed that, although the biotrophic phase is extended, the time to complete the asexual lifecycle on potato leaves is shorter compared to control isolates such as the Dutch 90128 (Cooke et al. 2012). The shorter latent period is thought to be the chief driver for the ability to outcompete other field isolates. In Figure 4, the drastic change in frequency of *P. infestans* MLGs is shown graphically for the period from 2003 to 2008. Also apparent is the rise in occurrence of

another MLG, 6_A1, which in 2011 became the most abundant isolate in the UK. It was thought that this switch coincided with higher temperatures during the 2011 growing season to which 6_A1 isolates are potentially better adapted (Cooke et al. 2012).

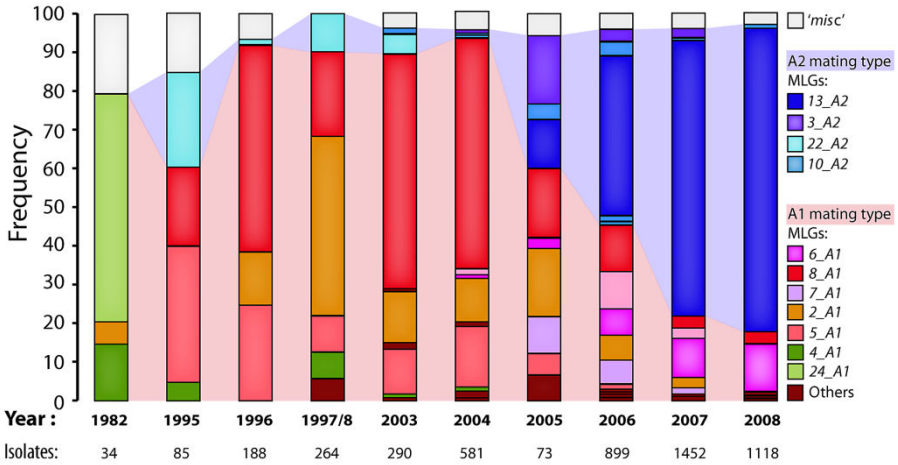


Figure 4 Frequency of different *P. infestans* field isolates in Great Britain from 2003 to 2008. Each colour represent a different genotype, and those depicted below the black line belong to mating type A1 and those above to A2 (reproduced from Cooke et al. 2012).

1.3.4 The *P. infestans* genome sequence reveals a large set of effector candidates

All identified *Phytophthora* avirulence genes so far belong to the group of RXLR-effector proteins, named after their highly conserved amino acid motif that enables the translocation of these proteins into the host cell (Haas et al. 2009;Whisson et al. 2007). These effectors share an N-terminal signal peptide, followed by an RXLR and dEER motif and a less conserved C-terminus responsible for the effector function. Another large effector group in *Phytophthora* species is the CRN- or "crinkler"-family, some of which induce crinkling and necrosis of the leaf tissue (Haas et al. 2009;Schornack et al. 2010). From the genome of the first sequenced *P. infestans* strain T30-4, 563 RXLR encoding genes and 196 CRN genes have been predicted (Haas et al. 2009). Recently, the release of the genome sequence of the MLG 13_A2 isolate 2006_3928A enabled genome wide comparisons to T30-4 and another sequenced isolate from the Netherlands referred to as NL07434 (Cooke et al. 2012). Several factors were identified that potentially contribute to the difference in pathogenicity, due to being either present and expressed or absent. One of the findings was that of all absent genes in 3928A, 21% encode RXLR effectors with homology to *Avr1*. In comparison to the other isolates, 3928A also

harbours six novel RXLR effector genes (Cooke et al. 2012). In total, 104 different RXLR effectors are expressed by 3928A during its lifecycle (Cooke et al. 2012). The three isolates share 45 RXLR genes, further termed the core effector set.

1.4 Breeding for disease resistance in potato

Wild species of *Solanum* form an important source for late blight resistance as many species have closely co-evolved with the late blight pathogen *P. infestans* in their natural habitat in South and Central America (Pel et al. 2009). These wild Section Petota *Solanum* accessions, many of which are maintained in the previously discussed wild potato collections, are considered to be a valuable source for new durable *resistance to P. infestans* (*Rpi)* genes. These genes can be introgressed into modern potato varieties via breeding programmes or deployed via genetic modification approaches (Haverkort et al. 2009).

1.4.1 Over 100 years of *R* gene introgression

The infamous late blight famine in Europe during the 1840s is believed to be the initiator of disease resistance breeding in crop plants (Hawkes 1993). The first reported attempts to incorporate resistance from wild species into cultivated *S. tuberosum* go as far back as 1909, when the British scientist Salaman crossed the hexaploid Mexican species *S. demissum* with *S. tuberosum* and retrieved *Phytophthora* resistant offspring (Salaman 1911). In the 1940s, 11 major *R* genes from *S. demissum* were identified and subsequently introgressed into *S. tuberosum* cultivars during the next 30 years (e.g. Gebhardt and Valkonen 2001; Umareus and Umareus 1994). Introgression of these single-gene resistances promised a good and efficient way for protecting crops (Golas et al. 2010). Unfortunately, often after the introduction of new cultivars harbouring one or even many of these *S. demissum R* genes into the field, the resistance was rapidly overcome by *P. infestans* (Turkensteen 1993). One example is the potato cultivar Pentland Dell, which had a stack of three R genes (*R1, R2* and *R3*) and had been released in 1960 and was commercially available in 1963. In 1967, only four years later, the resistances were overcome by *P. infestans* isolates (White and Shaw 2010). This resulted in a movement away from introgression of major *R* genes, towards breeding for quantitative trait loci (QTL) conferring non-race specific resistances or field resistance (Bradshaw et al. 2006;Sliwka 2004). However, recent growing seasons witnessed the breakdown of such strong field resistances in the cultivars ‘Stirling’ and ‘Lady Balfour’

(Cooke et al. 2012). During the 1990s the identification and molecular cloning of selected major *R* genes from wild potatoes came into the focus again. The chief driver was the identification of higher durability of some *Rpi* genes compared to *R1* to *R11*, and the idea of stacking several such *Rpi* genes derived from wild species (Tan et al. 2010).

1.4.2 Common approaches to identify genomic regions linked to a specific trait

Identification of resistance to a specific isolate of a pathogen is the first step towards determining genetic markers that co-segregate with the trait and establishing the physical position of the resistance. The mapping of *R* genes can be divided into several steps as described by e.g. Sliwka (2004): (i) crossing the resistant plant with a suitable susceptible clone; (ii) phenotyping the progeny for resistance or susceptibility, and bulking individuals with the same phenotype; (iii) creating a genetic map using molecular markers; and (iv) finding the markers that co-segregate with the resistance and position them on a genetic map.

Bulked segregant analysis (BSA) from step (ii) has since its description played a pivotal role in the identification of the chromosomal location of new *R* genes as well as other traits (Hormaza et al. 1994;Quarrie et al. 1999). This technique was first described by Michelmore et al. (1991) for lettuce populations that segregated for resistance to a mildew pathogen. The central idea for BSA is that two pools, or bulks, of plants differ in one trait, for example resistance or susceptibility to *P. infestans,* while all other traits are independently segregating. Markers that are found to be different between the bulks can then genetically be linked to the trait conferring loci (Michelmore et al. 1991). An advantage of this technique is the requirement for fewer individuals in a bulk (Michelmore et al. 1991) compared to analysing individuals from a segregating population.

For all previously mentioned *R* genes, three major genotyping techniques were used in the past, or are still in use: Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length Polymorphism (AFLP) or NBS profiling. RFLP utilises nucleotide changes that influence the enzymatic restriction digestion of DNA between members of a population. These polymorphisms can include a single nucleotide change, insertions, deletions or inversions (Sliwka 2004). Potato *Rpi* genes that were mapped using RFLPs are *R1, R6*, *R7* (el-Kharbotly et al. 1994), the *R3* locus (Huang et al. 2004), *RB* (Naess et al. 2000), *Rpi-mcq1.1*, *Rpi-mcq1.2* (Smilde et al. 2005), *Rpi-ber1* (Rauscher et al. 2006) and *Rpi-pnt1* (Kuhl et al. 2001). RFLPs were the genotyping method available when BSA was first described by Michelmore et al. (1991).

AFLPs also detect nucleotide polymorphisms. However these are manifested in PCR amplifications instead of enzymatic digestions. The *Rpi* genes *R2*, *R3a*, *R10* and *R11* were successfully mapped to their genetic position using AFLPs (Bradshaw et al. 2006;Huang et al. 2005;Li et al. 1998).

A more recent approach, first described by van der Linden et al. (2004), is the so called NBS profiling. This PCR-based approach targets the conserved NB-domain of NB-LRR type *R* genes and is designed for small segregating populations with bulks of three to ten individuals (Jacobs et al. 2010). PCR fragments that co-segregate with the resistance are sequenced and BLAST searches, in combination with literature information, are used to identify the map position. Two *R* genes have so far been mapped using NBS-profiling and include *Rpi-ver1* (Jacobs et al. 2010) and *R8* to chromosome 9 (Jo et al. 2011). *R8* was previously thought to be a member of the *R3* locus on chromosome 11 (Jo et al 2011).

To date, a total of 22 *Rpi* genes have been cloned from *Solanum* species (Jo et al. 2011;Vleeshouwers et al. 2011) and twenty more have been mapped to the potato chromosomes 4–11 with numbers increasing regularly (Table 1) (Hein et al. 2009a;Jo et al. 2011;Verzaux et al. 2011). All cloned *Rpi* genes belong to the class of CNLs, and are physically clustered together in groups of genes sharing high sequence similarity. *R* genes against other pests and diseases such as bacteria, nematodes, aphids, fungi and viruses are not discussed in this thesis as this is beyond the scope of this project.

In Table 1, the extent of mapped *Rpi* genes is presented in detail including the original references. One of the largest loci for functional *Rpi* genes is found on chromosome 4 where 13 *Rpi* genes were mapped and partially cloned, followed by chromosome 9 with 11 and chromosome 11 with seven *Rpi* genes (Table 1). Chromosome 8 contains 5 *Rpi* genes, and three are found on chromosome 10. So far, only one *Rpi* gene was mapped and cloned from chromosomes 5, 6 and 7 (Table 1).

Table 1 A large number of *Rpi* genes have been identified from wild potato species and have been mapped to various loci on potato chromosomes 4–11. These methods are based on linkage maps (including marker identification applying RFLP, AFLP or RAPD), mining for alleles of functional *Rpi* genes in different species using PCR amplification, or NBS profiling where degenerate primers are used to amplify a range of similar *Rpi* gene candidates.

Rpi gene	Strategy	*Solanum sp.*	Reference
chromosome 4			
R2	Map based	*S. demissum*	Li et al. 1998
R2-like	allele mining	*S. demissum*	Lokossou et al. 2009
Rpi-edn1.1	allele mining	*S. × edinense*	Lokossou et al. 2009
Rpi-snk1.1	allele mining	*S. schenckii*	Lokossou et al. 2009
Rpi-snk1.2	allele mining	*S. schenckii*	Lokossou et al. 2009
Rpi-hjt1.1	allele mining	*S. hjertingii*	Lokossou et al. 2009
Rpi-hjt1.2	allele mining	*S. hjertingii*	Lokossou et al. 2009
Rpi-hjt1.3	allele mining	*S. hjertingii*	Lokossou et al. 2009
Rpi-dmsf1	Map based	*S. demissum*	Hein et al. 2007
Rpi-blb3	allele mining	*S. bulbocastanum*	Lokossou et al. 2009
Rpi-abpt	allele mining	*S. bulbocastanum*	Lokossou et al. 2009
Rpi-mcd	Map based	*S. microdontum*	Sandbrink et al. 2000
Rpi-bst1	Map based	*S. brachistotrichum*	J. Jones and Z. Chu (communicated through Hein et al. 2009)
Rpi-mcd1	Map based	*S. microdontum*	Tan et al. 2008
chromosome 5			
R1	Map based	*S. demissum*	El Kharbotly et al. 1994;
chromosome 6			
Rpi-blb2	Map based	*S. bulbocastanum*	Vossen et al. 2003 + 2005
chromosome 7			
Rpi1	Map based	*S. pinnatisectum*	Kuhl et al. 2001
chromosome 8			
RB	Map based	*S. bulbocastanum*	Naess et al. 2000
Rpi-blb1	Map based	*S. bulbocastanum*	Park et al. 2005
Rpi-pta1	Map based	*S. papita*	Vleeshouwers et al. 2008
Rpi-sto1	Map based	*S. stoloniferum*	Vleeshouwers et al. 2009
Rpi-plt1	Map based	*S. polytrichon*	Wang et al. 2008
chromosome 9			
Rpi-dlc1	Map based	*S. dulcamara*	Golas et al. 2010
Rpi-vnt1.1	NBS profiling	*S. venturii*	Pel et al. 2009; Foster et al. 2009
Rpi-vnt1.2	NBS profiling	*S. venturii*	Pel et al. 2009; Foster et al. 2009
Rpi-vnt1.3	NBS profiling	*S. venturii*	Pel et al. 2009; Foster et al. 2009
Rpi-mcq1.1	Map based	*S. mochiquense*	Smilde et al. 2005
Rpi-mcq1.2	Map based	*S. mochiquense*	Smilde et al. 2005
Rpi-edn2	allele mining	*S. × edinense*	Nieks et al. 2011
Rpi-ver1	NBS profiling	*S. verrucosum*	Jacobs et al. 2010
R8	NBS profiling	*S.demissum*	Jo et al. 2011
R1-like	allele mining	*S. caripense*	Trognitz and Trognitz, 2005
Rpi-phu1	Map based	*S. phureja*	Sliwka et al. 2006
chromosome 10			
Rpi-ber1	Map based	*S. berthaultii*	Park et al. 2009
Rpi-ber2	Map based	*S. berthaultii*	Park et al. 2009
chromosome 11			
R3a	Map based	*S. demissum*	
R3b	Map based	*S. demissum*	Huang et al. 2004
R5	allele mining	*S. demissum*	Huang et al. 2005
R6	Map based	*S. demissum*	Huang et al. 2005
R7	Map based	*S. demissum*	El Kharbotly et al. 1996
R10	allele mining	*S. demissum*	El Kharbotly et al. 1996
R11	allele mining	*S. demissum*	Bradshaw et al. 2006
Rpi-pcs	Map based	*S. paucissectum*	Bradshaw et al. 2006
Rpi-avl1	Map based	*S. avilesii*	Villamon et al. 2005

1.4.3 *Rpi* and *Avr* gene pairs

For eight of the cloned *Rpi* genes the corresponding effectors have been identified and cloned. As these pathogen molecules trigger PCD and bring about avirulence upon *Rpi* genes based detection, they are also referred to as avirulence (*AVR*) genes, in contrast to the alleles that evade recognition which are termed virulence (Vir) genes. The cloned *P. infestans AVR* genes and their corresponding *Rpi* gene comprise: R1-AVR1, R2-AVR2, R3a-AVR3a, R3b-AVR3b, R4-AVR4, Rpi-blb1-AVRblb1, Rpi-blb2-AVRblb2, Rpi-vnt1-AVRvnt1 (reviewed in Vleeshouwers et al. 2011; Jo et al. 2011). With this knowledge, resistant potato accessions can be screened for recognition of these AVR genes or any other effector, which informs the presence of the corresponding *R* genes by eliciting a phenotypic PCD response. Thus, effectors provide a means to genetically track resistances and identify novel *Rpi* genes. These assays are usually carried out using *Agrobacterium tumefaciens* or virus-mediated transient expression (Bryan and Hein 2008;Vleeshouwers et al. 2011;Vleeshouwers et al. 2008). Sequencing the *P. infestans* genome and the discovery of its potential effector repertoire is a milestone towards large scale resistance tracking. Cloning and synthesis of effectors has resulted in several libraries that are being used to screen resistant potato plants for their resistance spectrum. The transient expression of 54 candidate effectors in a population segregating for late blight resistance enabled Vleeshouwers et al. (2008) to map and clone the corresponding *Rpi* gene. Another application is the co-expression of the effector and the cognate *R* gene in leaves of the model Solanaceae *Nicotiana benthamiana* to verify candidate *Rpi* genes (Vleeshouwers et al. 2008).

1.4.4 Sequencing based high-throughput mapping approaches

All above described methods for genotyping were successfully applied in identifying markers that are linked to a wide range of traits. However, they are labour intensive and have a low throughput when compared to recent sequencing-based genotyping methods (Miller et al. 2007). Developments in second-generation sequencing (SGS) methods and instrument platforms such as the Roche 454 for longer reads or the Illumina platforms for shorter reads, has greatly reduced the costs per nucleotide of sequence, by increasing the number of reads to thousands of million reads per instrument run (Gnirke et al. 2009;Metzker 2010). Not only has SGS led to an increasing number of whole genome sequences from a wide range of organisms, but also several approaches have been developed that exploit sequencing to discover a large number of SNP markers over entire genomes in large populations in a few steps (Davey et al. 2011).

A milestone for this development was the human genome project in which thousands of SNPs were identified over all chromosomes (Davey et al. 2011).

One example for genotyping technologies based on whole genome SNP information is the Illumina GoldenGate technology that is currently being exploited in potato and barley (Close et al. 2009). In the case of potato, a joint effort between Dr Robin Buell (MSU) and Dr Glenn Bryan (JHI) aided the discovery of DM related SNPs after alignment of SolCAP project derived EST contigs (http://solcap.msu.edu/) against the DM genome assembly. The EST contigs are derived from the potato cultivars Bintje, Kennebec and Shepody. After filtering for non-repetitive, high quality and non-overlapping SNPs, 1,920 unique and regularly spaced polymorphisms were retained that can be used to genotype large population sets (Dr Sanjeev K. Sharma, personal communication).

For many plant species whole genome sequences are not available yet. However second generation sequencing can still be used in genotyping when it is combined with random DNA restriction using specific enzymes. Examples for this technique comprise the so called restriction-site-associated DNA sequencing (RAD-seq) technology (Davey et al. 2011;Miller et al. 2007) and genotyping-by-sequencing (GBS; Elshire et al. 2011). RAD-seq, initially developed in combination with microarrays, has first been utilised with SGS in BSA for mutations in two model organisms, the fish threespine stickleback and the fungus *Neurospora crassa* (Baird et al. 2008). Similarly, GBS was successfully applied in genotyping and mapping of maize and barley populations (Elshire et al. 2011). BSA in conjunction with SGS was successfully used in fine-mapping a grain-protein content gene by Trick et al. (2012).

A new approach termed fast forward genetics applies BSA together with targeted genome enrichment and next-generation sequencing (Mokry et al. 2011). In a proof of principle study, Mokry and colleagues were able to identify a novel factor for stem cell activity from *Arabidopsis* mutant lines. In this approach, the target region was first mapped by using “light sequencing” data of up to 10× genome coverage. Subsequently, based on the identified chromosomal region, a capture array was designed and applied to re-sequence enriched DNA samples for the specific target region and identify SNPs using statistical methods (Mokry et al. 2011). This is an example of how sequencing targeted regions can lead to a more efficient data output by increasing the sequencing depth (Gnirke et al. 2009;Mamanova et al. 2010;Stitziel et al. 2011).

Currently, four methods are available for capture and enrichment of targeted genomic regions and have been compared by Mamanova et al. (2010). The methods are PCR,

Molecular Inversion Probes and the in-solution and on array hybrid capture technologies. The latter two are the most widely used capture technologies.

1.5 Scope of this Thesis

Plants possess an innate immune system that is inherited and originally derived from the imminent threat of pests and pathogens. Key-players in this disease resistance are NB-LRR protein encoding genes that directly or indirectly perceive small pathogen derived molecules and trigger a hypersensitive response in form of a local cell death. Potato breeding strategies focus next to quality traits also on the introduction of novel resistances from wild relatives. However the identification and characterisation of the underlying resistances is a time-consuming task and novel strategies are needed. The aim of this thesis is to characterise the NB-LRR gene complement of potato and utilise this in combination with second generation sequencing technologies to accelerate the identification of new *Rpi* genes from wild potato species. Chapter 2 focuses on the identification and characterisation of the NB-LRR gene family from the sequenced potato clone DM. These analyses also comprise establishing the phylogenetic relationships between members of the NB-LRR gene family as well as their physical position on the 12 potato chromosomes.

Chapter 3 elaborates on the utilisation of this NB-LRR sequence information in the design of a NB-LRR gene specific sequence capture tool. A proof-of-concept study on the sequenced potato clone DM shows the enrichment efficiency as well as how this tool can be used to further annotate the NB-LRR gene family in DM and potentially in other Solanaceae.

Chapter 4 focuses on the extraordinary late blight resistance of the potato cultivar Sarpo Mira, and how BSA and genotyping were applied to shed light on this resistance.

The general findings of this PhD, the implication for future research and *Rpi* gene cloning are discussed in the final Chapters 5 and 6.

Chapter 2 - Identification and localisation of the NB-LRR gene family within the potato genome

The study presented in this chapter resulted in a publication in BMC Genomics (Jupe et al 2012). The data were produced with the help of several people from the James Hutton Institute in Dundee and The Sainsbury Laboratory in Norwich. Work that was carried out by others has been marked clearly in the results and material and methods section. Whole sections in this chapter are direct statements from Jupe et al. (2012).

2.1 Introduction

As introduced in Chapter 1, potato plants face a constant barrage of pest and microbial threats. More than 50 functional NB-LRR genes have been cloned from potato and related members of the Solanaceae (Hein et al. 2009a), however many resistances have already been broken by contemporary isolates and forms of potato pests and diseases. To be able to better understand the co-evolution between the potato and its threats, as well as the genetic resistance potential, the full resistance gene complement must be known. Previously, 738 NB-LRR-like sequences have been identified in a BAC library prepared from a heterozygous diploid potato clone, RH (Bakker et al. 2011). The genome sequence of the doubled monoploid *Solanum tuberosum* group Phureja clone DM1-3 516 R44 (hereafter referred to as DM), has recently been described (PGSC 2011). Among the 39,031 annotated protein coding genes, 408 NB-LRR coding genes were predicted but not further characterised. In this study, we used a process of iterated computational and manual annotation to further identify potential NB-LRR coding sequences, determine their locations on the 12 potato chromosomes and study the phylogenetic and positional relationships between the individual genes. Our results provide significant insight into the evolution of NB-LRRs and, importantly, a blueprint for future efforts to identify and more rapidly clone functional NB-LRR genes from *Solanum* species.

2.2 Results

2.2.1 Identification of NB-LRR genes within the DM genome protein models

All protein classes share highly conserved amino acid motifs. These motifs can thus be used for the identification of novel candidates from un-annotated sequence data. The motif-based sequence analysis tool MEME (Bailey and Elkan 1994) was used in conjunction with a positive sequence set of 53 characterised NB-LRR protein sequences from diverse plant species and a negative sequence set containing diverse nucleotide binding protein and PRR sequences (Annex 1) to identify 20 sequence motifs putatively characteristic of NB-LRR proteins. Some of the disclosed motifs (Table 2) are associated with known domains from the TNL and CNL subfamilies. Of those, 13 encompass previously described features of the NB-LRR family, such as the P-loop, RNBS-A non-TIR, RNBS-B, RNBS-C, RNBS-D, GLPL, LRR-motif 1 (LDL), MHDV, TIR-1, TIR-2, TIR-3 (Meyers et al. 1999), EDVID (Rairdan et al. 2008), and Kin-2 (Tarr and Alexander 2009) domains.

The 20 potentially characteristic motifs were used as queries in a search using the motif alignment and search tool (MAST) (Bailey and Gribskov 1998) against a combination of the annotated potato genome v3.4 DM protein models (DMP) and the training set sequences used to derive the motifs. In total, 765 DMPs were identified that contained the motifs identified by MEME, with an E-value of less than 2 (Figure 5). The positive and negative training set sequences could be distinguished with 100% specificity on the basis of reported E-values. In total, 343 DMP sequences had reported E-values less than the highest seen for a member of the positive training set ($E < 2.7e-45$). A further 134 DMP sequences had E-values less than the smallest E-value observed for a member of the negative training set ($E < 8.5e-24$). Thus, a total of 477 candidate NB-LRR DMP sequences were identified on the basis of motif composition (Figure 5). These results were achieved with help of Dr Graham Etherington (TSL) and Dr Leighton Pritchard (JHI).

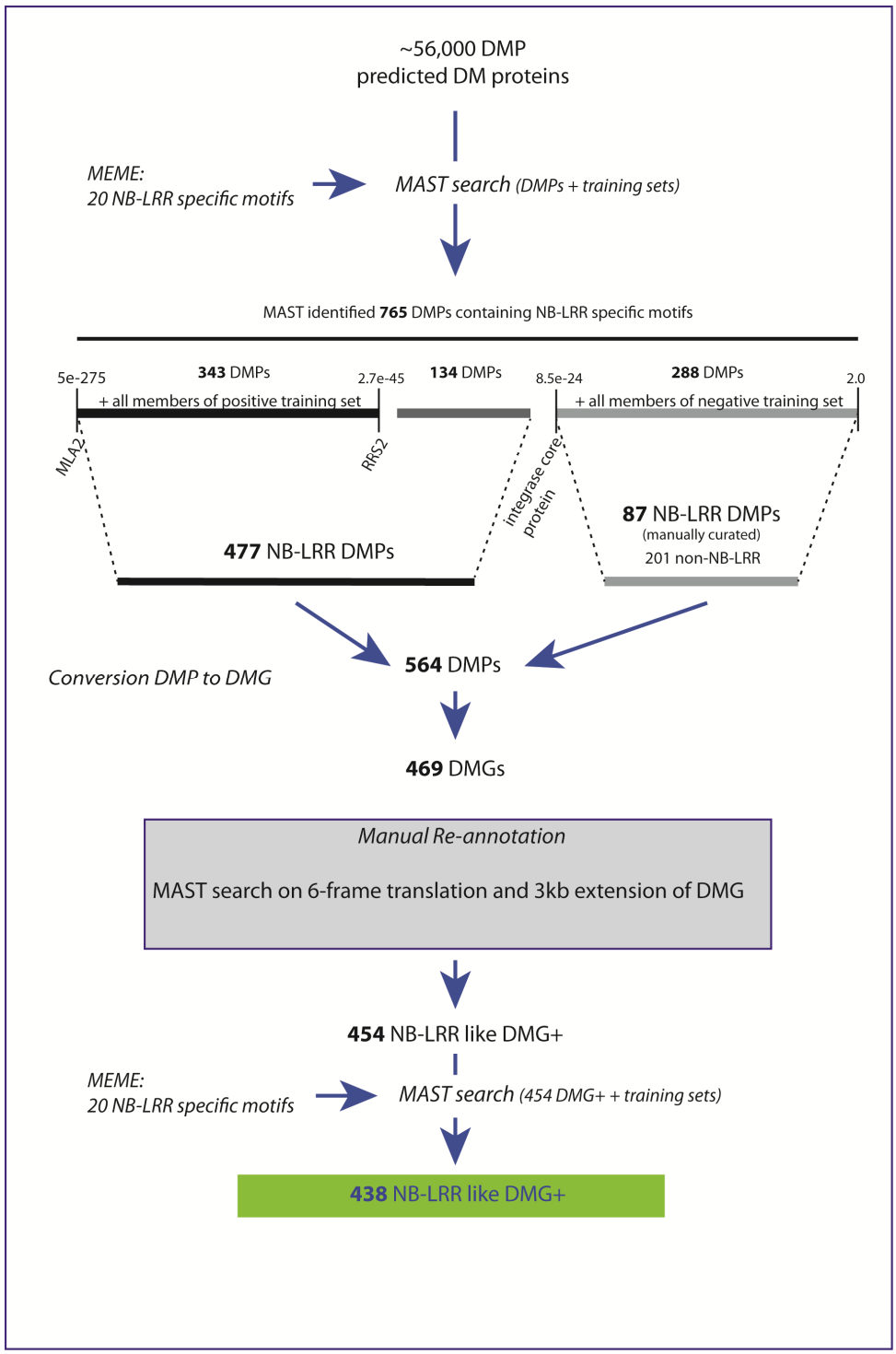


Figure 5 Graphical overview of the MAST search output ranked according to the E-value scores obtained for MEME motifs. By including DMPs that yielded an E-value score of up to 2.0, 765 proteins were identified. Within the E-value range of the negative training set, 87 sequences encoded for very short DMPs and contained additional NB-LRR gene associated domains in the extended DMP+ sequence.

Table 2 NB-LRR-specific amino acid motifs identified with psp-gen script MEME (Bailey et al. 2010). a) Motifs are listed according to their ranking derived from the psp-gen MEME analysis. b) Consensus amino acid sequence derived from psp-gen MEME analysis. References for known motifs encompassed in the MEME motifs are shown.

Motif[a]	Sequence[b]	Domain	Group	similar to	Reference
motif 1	PIWGMGGVGKTTLARAVYNDP	NB-ARC	CNL/TNL	P-loop	Meyers et al. (1999)
motif 2	LKPCFLYCAIFPEDYMIDKNKLIWLWMAE	NB-ARC	CNL	RNBS-D	Meyers et al. (1999)
motif 3	CGGLPLAIKVWGGMLAGKQKT	NB-ARC	CNL/TNL	GLPL	Meyers et al. (1999)
motif 4	YLVVLDDVWDTDQWD	NB-ARC	CNL/TNL	Kin-2	Tarr and Alexander (2009)
motif 5	NGSRIIITTRNKHVANYMCT	NB-ARC	CNL/TNL	RNBS-B	Meyers et al. (1999)
motif 6	HFDCRAWVCVSQQYDMKKVLRDIIQQVGG	NB-ARC	CNL	RNBS-A	Meyers et al. (1999)
motif 7	CRMHDMMHDMCWYKAREQNFV	linker	CNL/TNL	MHDV	Meyers et al. (1999)
motif 8	MEDVGEYYFNELINRSMFQPI	linker	CNL/TNL	-	
motif 9	LIHLRYLNLSGTNIKQLPASI	LRR1	CNL/TNL	Motif1 LDL	Meyers et al. (1999)
motif 10	LSHEESWQLFHQHAF	NB-ARC	CNL/TNL	RNBS-C	Meyers et al. (1999)
motif 11	MPNLETLDIHNCPNLEEIP	LRR	CNL/TNL	-	
motif 12	IMPVLRLSYHHLPYH	NB-ARC	CNL/TNL	-	
motif 13	QIVIPIFYDVDPSDVRHQTGSFGEAFWKHCSR	TIR	TNL	TIR-3	Meyers et al. (1999)
motif 14	AIKDIQEQLQKVADRRDRNKVFVPHPTRPIAIDPCLRALYAEATELVGIY	monocot	-	-	
motif 15	KNYATSRWCLNELVKIMECKE	TIR	TNL	TIR-2	Meyers et al. (1999)
motif 16	DAAYDAEDVIDSFKYHA	pre-NB	CNL	EDVID	Rairdan et al. (2008)
motif 17	FAIPKLGDFLTQEYYLHKGIKKEIEWLKRELEFMQA	pre-NB	CNL	-	
motif 18	KYDVFLSFRGADTRRTFTSHLYEALKNRGINTF	TIR	TNL	TIR-1	Meyers et al. (1999)
motif 19	IKMVEITGYRGTRFPNWMGHPVYCNMVSISIRNCKNCSCLP	LRR	CNL/TNL	-	
motif 20	ETSSFELMDLLGERWVPPVHLREFKSFMPSQLSALRGWIQRDPSHLSNLS	monocot	-	-	

2.2.2 Manual re-annotation of DM gene models containing NB-LRR-like sequences

Manual inspection of the remaining 288 DMPs whose E-values lay above the 8.5e-24 cut-off indicated that several sequences contained motif patterns potentially characteristic of NB-LRR proteins, but that were truncated or otherwise distorted (Figure 4). Of these, 87 sequences that contained at least two TIR/CC-specific motifs, or three NB-ARC specific motifs, were noted as potential errors in automated gene calling or annotation and carried forward into the candidate set pending a manual check, to give a total of 564 putative NB-LRR DMP sequences (Figure 5).

Several of the candidate DMP sequences were derived from the same DM gene model (DMG) sequence as alternative transcripts. We found that 469 distinct DMG sequences coded for the 564 candidate NB-LRR sequences (Figure 5). The MAST search was repeated against conceptual translations of these 469 DMGs, and indicated that 277 DMG translations apparently lacked domains characteristically associated with TNL or CNL genes. To investigate if mis-annotation might be responsible for this, these DMG sequences were extended by 3kb at both the 5' and 3' ends to generate a counterpart DMG+ sequence set. The MAST search was repeated against the conceptual translations of the DMG+ sequences. We found that all 277 DMG sequences that initially lacked

typical NB-LRR domains contained additional MEME motifs in an order characteristic of the other candidate NB-LRR sequences.

Gene models corresponding to the DMG+ sequences were modified to incorporate the additional characteristic motifs identified above. Conceptual translations of these genes (referred to as DMP+ sequences), were compared to NB-LRR proteins in the nr database at NCBI using BLASTP (Altschul et al. 1990) to identify potential introns and start and stop codons. In addition, six DMG+ models appeared to encode two complete NB-LRR-like sequences each, and were split into a total of twelve distinct gene models. A further 15 NB-LRR-like sequences appeared to have been split across two adjacent DMGs in the initial annotation. Thus, the number of identified NB-LRR-like sequences after manual correction was 454 (Figure 5). A further MAST search was carried out on these sequences, from which 438 DMG sequences were found to have an E-value less than that for any member of the negative sequence set (Digital accompanying material E1, Figure 5). Sequences are enclosed in Digital accompanying materials E4 and E5.

In total, 154 of the predicted NB-LRR sequences are encoded by a single reading frame without introns. A further 110 predicted NB-LRRs contain a single intron and/or a frameshift, and 100 genes contain two introns and/or frameshifts. The remaining 74 genes have between three and eight introns and/or frameshifts (Digital accompanying material E1). Without further detailed analysis (e.g. RNA sequencing), it is difficult to determine if the predicted introns and/or frameshifts are genuine or a result of sequencing/assembly errors. However, of the 154 candidate NB-LRR genes without an intron, 116 contain all domains associated with TNLs or CNLs and are thus referred to as ‘full length’. A further 97 genes that contain one or two potential introns but no frameshift are also classified as ‘full length’ on the same grounds. Among the other DMG+ sequences, 155 contain all domains associated with TNLs or CNLs, and are labelled as ‘potentially full length’. The remaining 70 genes are classified as ‘partial’, as they show truncations within the N-terminal domains and/or absence of LRR domains. The average length of the coding sequence for partial genes is 1kb, for full length and potentially full length genes 3kb, and for all identified NB-LRR genes combined 2.7kb (Digital accompanying material E1).

Based on the presence of the TIR domain derived motifs (Table 2 motifs 13, 15 and/or 18), 77 genes were identified as TNLs. These data were verified using a Pfam (Bateman et al. 2002) search over all sequences. All 55 full length and potentially full length TNLs share the TNL discriminating aspartic acid (D) in the final position of the Kin-2 domain (Cannon et al. 2002;Meyers et al. 1999;Tarr and Alexander 2009). The 316 (potentially) full length non-TIR sequences encode for a tryptophan (W) in this position, and contain

the CNL specific motifs 16 and/or 17 (Table 2). This analysis was further corroborated by the presence of the CNL-type NB-ARC motifs 2 and 6 (Table 2), that encapsulate RNBS-D and RNBS-A, described by Meyers et al. (1999). A Paircoil2 analysis (McDonnell et al. 2006) was carried out with the help of Dr Graham Etherington, on the positive training set (Annex 1) to establish the conditions for coiled-coil domain predictions in well annotated genes. The highest minimum p-score for a functional CC-NB-LRR gene was found for *Rpi-vnt1* (Foster et al. 2009) with 0.047 starting at amino acid position 73. The latest start position of a CC domain was determined for *R2* and *Rpi-blb3* at amino acid position 98. To determine the presence of CC motifs within the 438 predicted NB-LRRs, a p-score cut-off of 0.047 was used for domains starting within the first 98 amino acids. Under these conditions, 107 NB-LRR genes were identified that contain a predicted CC domain. A total of 254 CNL genes do not contain a predicted CC domain. The TNL and CNL predictions are summarised in Table 3 and compared to the initial analysis from the PGSC analysis (PGSC 2011). Amongst the predicted TNLs and CNLs, homologues of the functionally characterised Solanaceae *R* genes *Gpa2, NRC1, R1, R2, Rpi-bt1*, *Rpi-blb2, Rpi-blb3, Rpi-vnt1*, and *Rx* were identified with more than 80% sequence identity using BLASTP. Further homologues of other functionally described Solanaceae *R* genes were identified, though with lower percentage sequence identity (Table 4).

Table 3 Comparison between DM NB-LRR genes identified and re-annotated in this study (left) with the data published by the Potato Genome Sequencing Consortium (right) (PGSC 2011). Partial genes (TIR-NB, CC-NB, NB-ARC) and (potential) full length genes (TIR-NB-LRR, CC-NB-LRR, NB-LRR) are shown. # represents numbers.

	NB-LRRs		PGSC	
	#	%	#	%
TNL	**77**	**17.6**	**49**	**12.0**
TIR-NB	22	5.0	14	3.4
TIR-NB-LRR	55	12.6	35	8.6
CNL	**361**	**82.4**	**359**	**88.0**
CC-NB	4	0.9	22	5.4
CC-NB-LRR	103	23.5	60	14.7
NB-LRR	213	48.6	172	42.2
NB-ARC	41	9.4	105	25.7
total	**438**		**408**	

Table 4 Comparison of functionally characterised Solanaceae *R* genes to DM NB-LRR cds. E-values, pairwise identity and coverage were established using BLASTP. The chromosome and cluster positions are shown alongside the phylogenetic group information.

Query	Organism	DM NB-LRR hit	E Value	% Pairwise Identity	Query coverage	Phylogenetic group	Chromosome	Cluster
NRC1	*N. tabacum*	PGSC0003DMG401026043	0	93.00%	99.32%	-	1	C3
Rx	*S. tuberosum*	PGSC0003DMG402007871	0	89.40%	96.05%	CNL-2	12	C56
Gpa2	*S. tuberosum*	PGSC0003DMG400007867	0	87.70%	99.01%	CNL-2	12	C56
PSH-RGH6	*S. tuberosum*	PGSC0003DMG402007871	0	87.00%	99.31%	CNL-2	12	C56
Rpi-vnt1	*S. venturii*	PGSC0003DMG401020585	0	86.50%	94.36%	CNL-4	9	C42
Rpi-blb2	*S. bulbocastanum*	PGSC0003DMG400021986	0	85.00%	100.00%	CNL-1	6	C28
Rpi-bt1	*S. bulbocastanum*	PGSC0003DMG402021043	0	84.40%	100.00%	CNL-6	unmapped	-
Rpi-blb3	*S. bulbocastanum*	PGSC0003DMG400011920	0	83.20%	99.65%	CNL-5	unmapped	-
R2	*S. demissum*	PGSC0003DMG400032572	0	81.60%	99.88%	CNL-5	4	C12
R1	*S. demissum*	PGSC0003DMG400003380	0	80.70%	97.68%	-	5	C24
TVR-A/B	*S. lycopersicum*	PGSC0003DMG400011898	0	79.80%	100.00%	-	9	C43
Gro1-4	*S. tuberosum*	PGSC0003DMG400017317	0	79.20%	100.00%	TNL	7	singleton
Mi1.2	*S. lycopersicum*	PGSC0003DMG400021986	0	79.00%	100.00%	CNL-1	6	C28
Rpi-blb1/RB	*S. bulbocastanum*	PGSC0003DMG400030855	0	78.80%	99.79%	CNL-6	8	C40
R3a	*S. demissum*	PGSC0003DMG401018576	0	78.70%	100.00%	CNL-8	11	C53
Tm-2	*S. tuberosum*	PGSC0003DMG400020584	0	77.60%	100.00%	CNL-4	9	C42
NL25	*S. lycopersicum*	PGSC0003DMG402002428	0	76.40%	96.25%	TNL	6	C31
I2	*S. lycopersicum*	PGSC0003DMG401018576	0	75.40%	100.00%	CNL-8	11	C53
Bs4	*S. lycopersicum*	PGSC0003DMG400018428	0	74.30%	98.50%	TNL	5	C22
Hero	*S. lycopersicum*	PGSC0003DMG400029504	0	71.00%	100.00%	-	4	C10
RGA2-Ca	*C. annuum*	PGSC0003DMG400009324	0	69.00%	99.69%	CNL-6	8	C41
ry-1	*S. tuberosum*	PGSC0003DMG400015681	0	65.60%	92.90%	TNL	11	C49
Bs2	*C. annuum*	PGSC0003DMG402002917	3.40E-178	64.50%	55.00%	-	12	C57
NL27	*S. lycopersicum*	PGSC0003DMG402016979	0	63.30%	88.31%	TNL	11	C49
N	*N. tabacum*	PGSC0003DMG400018428	0	54.60%	98.42%	TNL	5	C22

2.2.3 Phylogenetic analysis

To study the evolutionary relationships among the predicted NB-LRR genes, a phylogenetic tree was estimated from the protein alignment of the conserved NB-ARC domains. This analysis was carried out together with Drs Katrin MacKenzie and Frank Wright (Biomathematics and Statistics Scotland, Dundee). Predicted NB-LRR genes containing ambiguous nucleotides in the NB-ARC domain were removed prior to the alignment. In addition to 413 predicted TNLs and CNLs, 33 functional NB-LRR genes from the positive training set were also included in the analysis. As expected (e.g. Meyers et al. 1999), the phylogenetic analysis separates the TNL and CNL gene products into two distinct clades and confirms thus our TIR motif prediction above (Figure 6 and Digital accompanying material E2). The TNL clade contains 68 NB-LRR sequences of which 6 are partial, missing motifs 2 and 6 (Table 2). The 68 NB-LRR can be divided into six small subgroups. Physical mapping of these (Figure 7 and Digital accompanying material E7) indicates that members of five subgroups are distributed over several chromosomes (Figures 2.2 and 2.3). Only members of one subgroup reside predominantly (8 out of 9) in a NB-LRR gene cluster on chromosome 9 (Figure 7 and Digital accompanying material E7).

Only a single DMG product, PGSC0003DMG400007999 (DMG identifiers hereafter are shortened to the last seven informative digits; DMG 0007999), could not reliably be placed in either of the CNL or TNL clades. The encoded gene product shows high sequence similarity (including the conserved TVS and PKAE amino acid motifs) to the atypical *Arabidopsis*/potato ADR1 CC-NB-LRR protein (Chini and Loake 2005).

Bootstrap support is given that further divides the CNL clade into CC_{RPW8}-type sequences (referred to as CNL-R) (Collier and Moffett 2009), and the canonical CNL proteins, that, with the exception of DMGs 0029313, 0029314 and 0029405, contain the EDVID motif (CC_{EDVID}-type) which is typically associated with CNLs (Rairdan et al. 2008). The CNL branch contains eight highly conserved subgroups (CNL-1 to CNL-8) amongst more diverse sequences and subgroups. CNL-1 contains 18 genes that map, with one exception, to chromosome 6. Members of this subgroup are homologous to the functional resistance to *Phytophthora infestans (Rpi)* genes *Rpi-blb2* (van der Vossen et al. 2005) and *Mi-1* (Milligan et al. 1998). CNL-2 members show sequence similarity to the functionally validated genes *Gpa2* and *Rx* (Bendahmane et al. 2000).

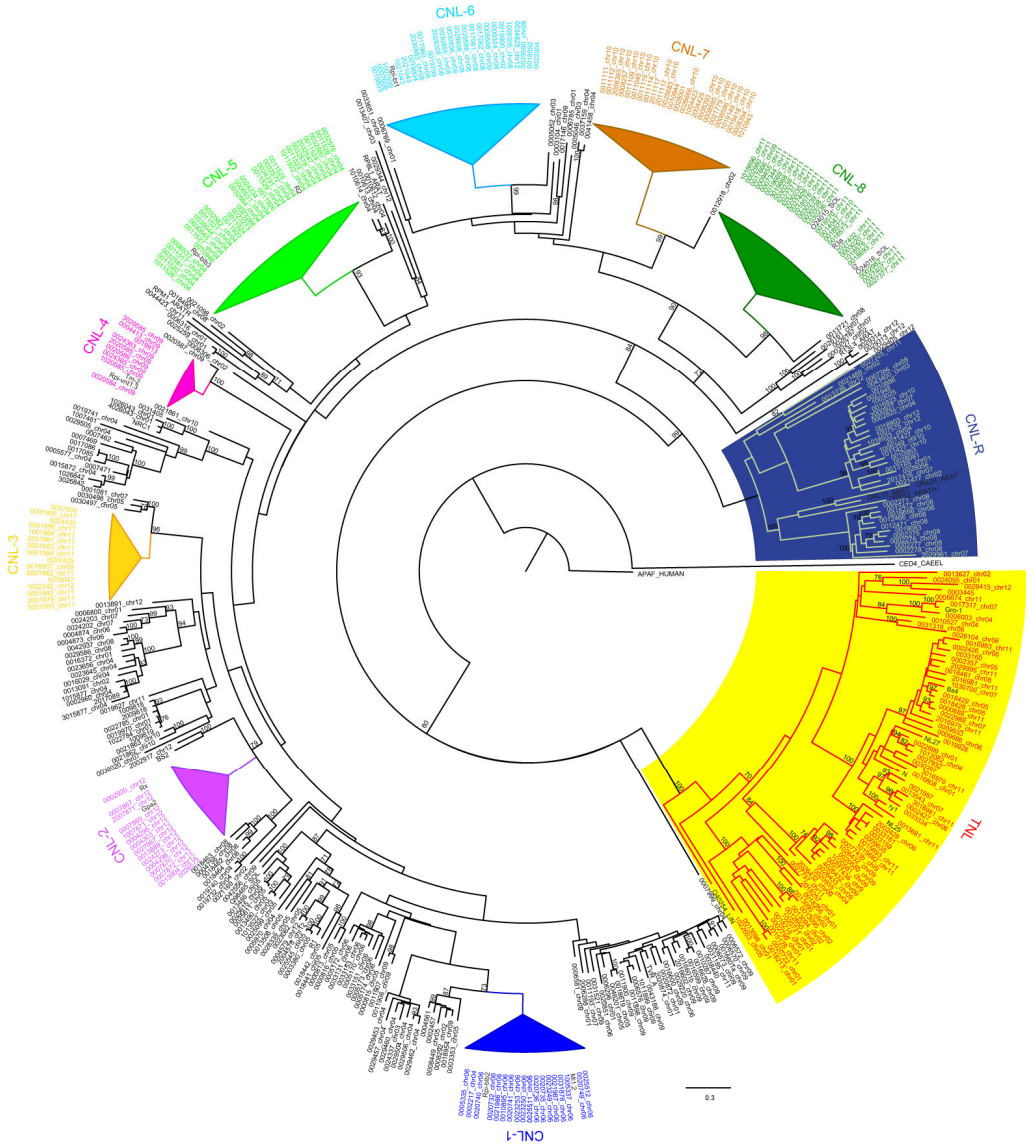


Figure 6 Maximum Likelihood Phylogenetic analyses of the predicted DM NB-LRR genes. The NB-ARC domains of TNL and CNL type genes were used, alongside selected NB-ARC domains from functional resistance genes, to study the phylogenetic relationships between them. Subgroups with highly similar gene products are marked: TNL genes have a yellow background, CNL-R type NB-LRR genes a blue background and CNL-1 to CNL-8 are shown in various colours. The gene product labels contain the 7 last informative digits from the DMG identifier, followed by their chromosomal position if known. Identifiers in black within coloured groups represent functionally characterized *R* genes. Percentages for bootstrap trees also resolving that clade are shown when over 70 %. The scale bar represents 30% divergence.

Apart from one gene, which resides on a yet unanchored superscaffold, the remaining 14 members reside on chromosome 12. The subgroup CNL-3 contains 16 members, of which four remain on unanchored superscaffolds. There is a single gene from this subgroup located on each of chromosomes 9 and 12, and ten genes on chromosome 11. Members of the smallest subgroup CNL-4 are homologous to *Rpi-vnt1* (Foster et al.

2009;Pel et al. 2009) and *Tm-2* (Lanfermeijer et al. 2003). The eight mapped members reside on chromosome 9 and one gene remains unmapped. The largest subgroup, CNL-5, contains 30 genes of which six remain unmapped and 24 reside on chromosome 4. Functionally validated *R* genes with sequence similarity to this subgroup include *R2* and *Rpi-blb3* (Lokossou et al. 2009;Park et al. 2005). Half of the 24 members of CNL-6 map to chromosome 8, one each to chromosome 2, 9 and 12 respectively, and the remaining nine are unmapped. The *Rpi-blb1/RB* (Song et al. 2003;van der Vossen et al. 2003) and *Rpi-bt1* (Oosumi et al. 2009) genes share sequence similarity with this group. Of the 24 sequences in CNL-7, 17 are localised on chromosome 10, one on chromosome 4 and six did not map to any of the chromosomes in this assembly. The CNL-8 subgroup contains 26 sequences. The physical mapping of these genes has placed 24 on chromosome 11 and the remaining two on chromosomes 9 and 10. The functionally validated potato and tomato *R* genes *R3a* (Huang et al. 2005), *R3b* (Li et al. 2011) and *I2* (Ori et al. 1997) share sequence similarities with members of this group.

2.2.4 NB-LRR gene mapping and physical clustering

Physical map positions were established for 370 (84%) of the annotated NB-LRR genes, on the 12 pseudomolecules described of the publicly available potato genome v3_2.1.10 (see Materials and Methods) and visualised using Biopython with the help of Dr Peter Cock (Cock et al. 2009) (Figure 7 and Digital accompanying material E7). CNLs are present on all 12 chromosomes whilst TNLs are absent from chromosomes 3 and 10 (Figures 2.3 and 2.4). The greatest number of NB-LRRs is found on chromosomes 4 and 11, harbouring 57 and 54 genes, respectively. Chromosome 3 contains the smallest number of NB-LRR genes (four) (Figure 8). From the map positions, NB-LRR gene clusters were determined by a combination of two previously described approaches (Meyers et al. 2003;Yang et al. 2008b). To form a cluster, the distance between neighbouring NB-LRRs was required to be less than 200kb, and fewer than eight non-NB-LRR genes between TNLs or CNLs. This approach identifies 63 clusters containing a total of 271 NB-LRRs (Figure 8). Thus 27% of the mapped NB-LRR genes appear not to be organised in physical clusters. Of the 63 clusters, 50 (79%) are homogeneous in that they contain only predicted NB-LRRs with a recent common ancestor, whereas the remaining clusters are heterogeneous, as they contain more distantly-related NB-LRRs.

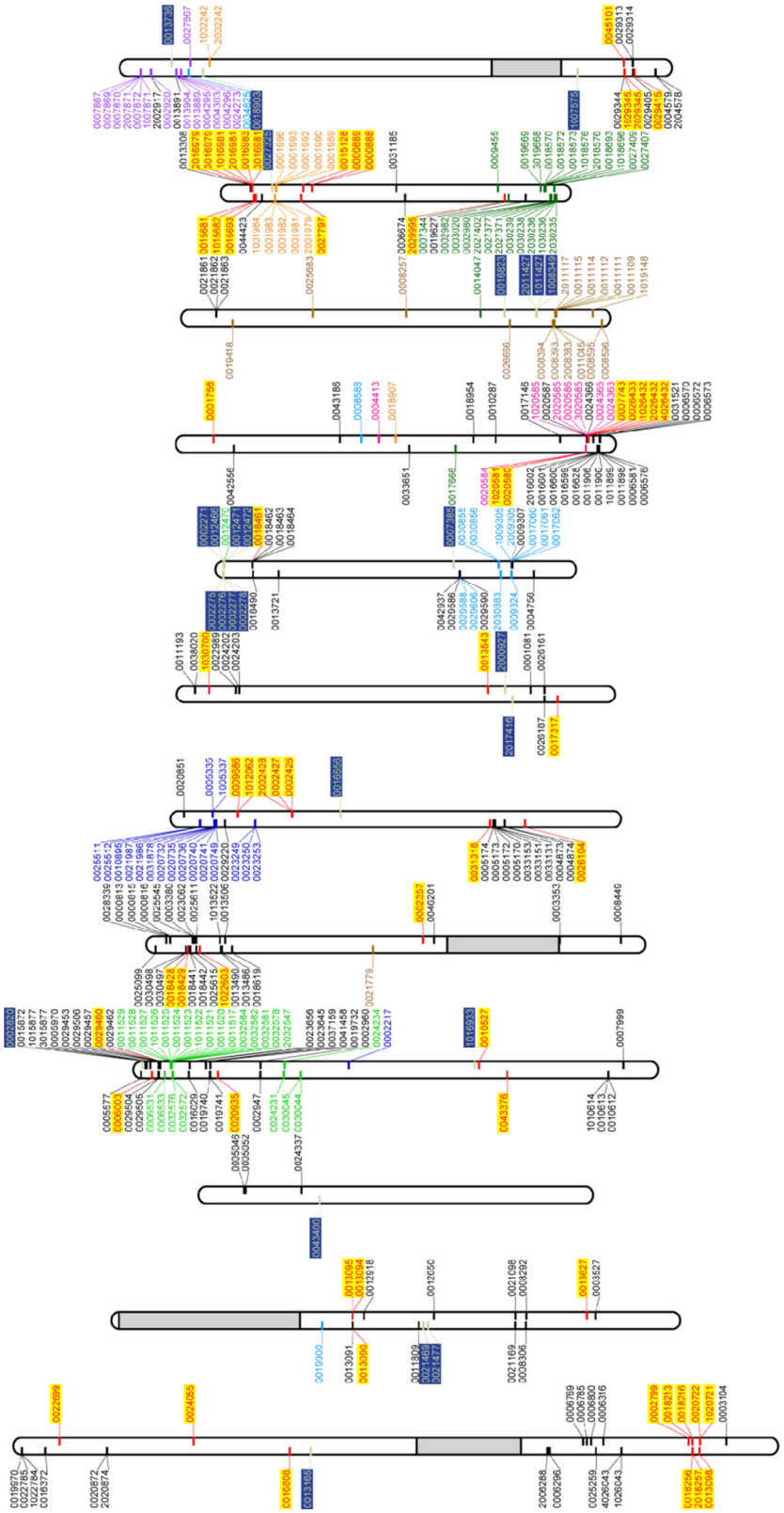


Figure 7 Physical maps of the 12 potato chromosomes with individual CNLs and TNLs. The relative map position of 366 unique DMGs encoding for NB-LRR type genes is shown on the individual pseudomolecules depicting the chromosomes 1–12. Each gene has a unique label representing the 7 last informative digits from the DMG identifier. Genes encoded by the positive DNA strand are depicted on the left hand side of the chromosomes, whereas those encoded by the negative strand are shown on the right. Colours and background of the genes are identical to the phylogenetic subgroups (TNL, CNL-R, CNL-1 to CNL-8) shown in Figure 1. Grey bars on chromosomes 1, 2, 5 and 12 represent known gaps in the assembly.

Chromosome 4 contains the greatest number of NB-LRR genes (57) and also the largest number of clusters (11). With the exception of cluster C10, which contains five homologues of the *R* gene *Hero* and one TNL, all remaining clusters on this chromosome are homogeneous clusters. The sizes of the clusters vary between two and 18 NB-LRR genes (Digital accompanying material E7). Eleven genes on chromosome 4 are not organised in clusters. The physically expanded and well described *R2* and *Rpi-blb3* locus (Lokossou et al. 2009) is located on this chromosome and its DM homologues are organised in the phylogenetic subgroup CNL-5 which spans four physical clusters (Figure 9a). Eighteen members form the homogeneous cluster C12, which is also the largest of all. The remaining members of CNL-5 are found in cluster C11, and two more are grouped (in C17 and C18) downstream of the bulk of the clusters.

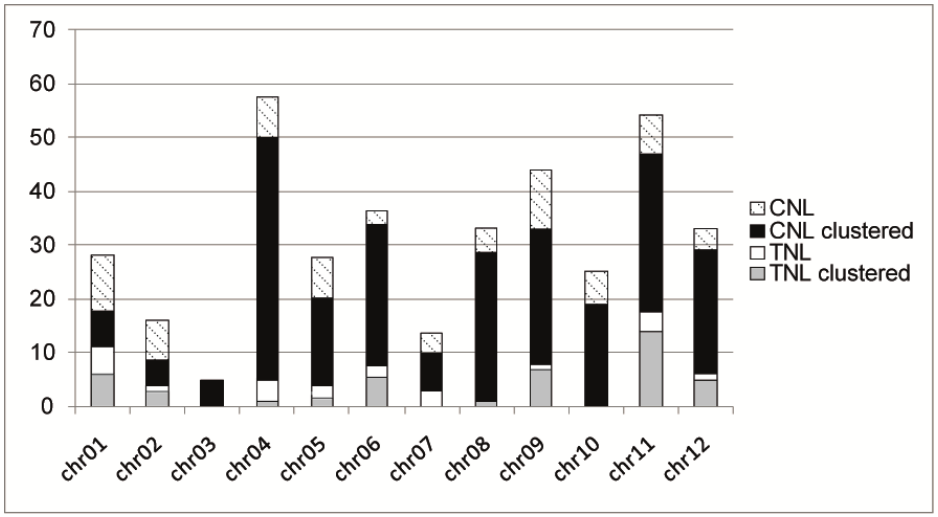


Figure 8 CNL and TNL organization within the potato genome. The distribution of NB-LRR genes is shown for each chromosome. Bars are divided into CNL genes (white-textured for non-clustered genes and black for those found in clusters) and TNL genes (white for non-clustered genes and grey for clustered TNLs).

The heterogeneous *R3* locus that contains the *Rpi* genes *R3a* (Huang et al. 2005) and *R3b* (Li et al. 2011) resides on the distal end of the long arm of chromosome 11. As mentioned, DM homologues of *R3a* and *R3b* form the phylogenetic subgroup CNL-8. Of the 26 members in this subgroup, 24 map to chromosome 11. *R3a* homologues are organised in three neighbouring homogeneous clusters: C52, C53 and C54 that contain two, seven and four members respectively. Two additional single *R3a* homologues are located upstream of C52. *R3b* homologues are organised in cluster C55 which harbours nine members (Figure 9b).

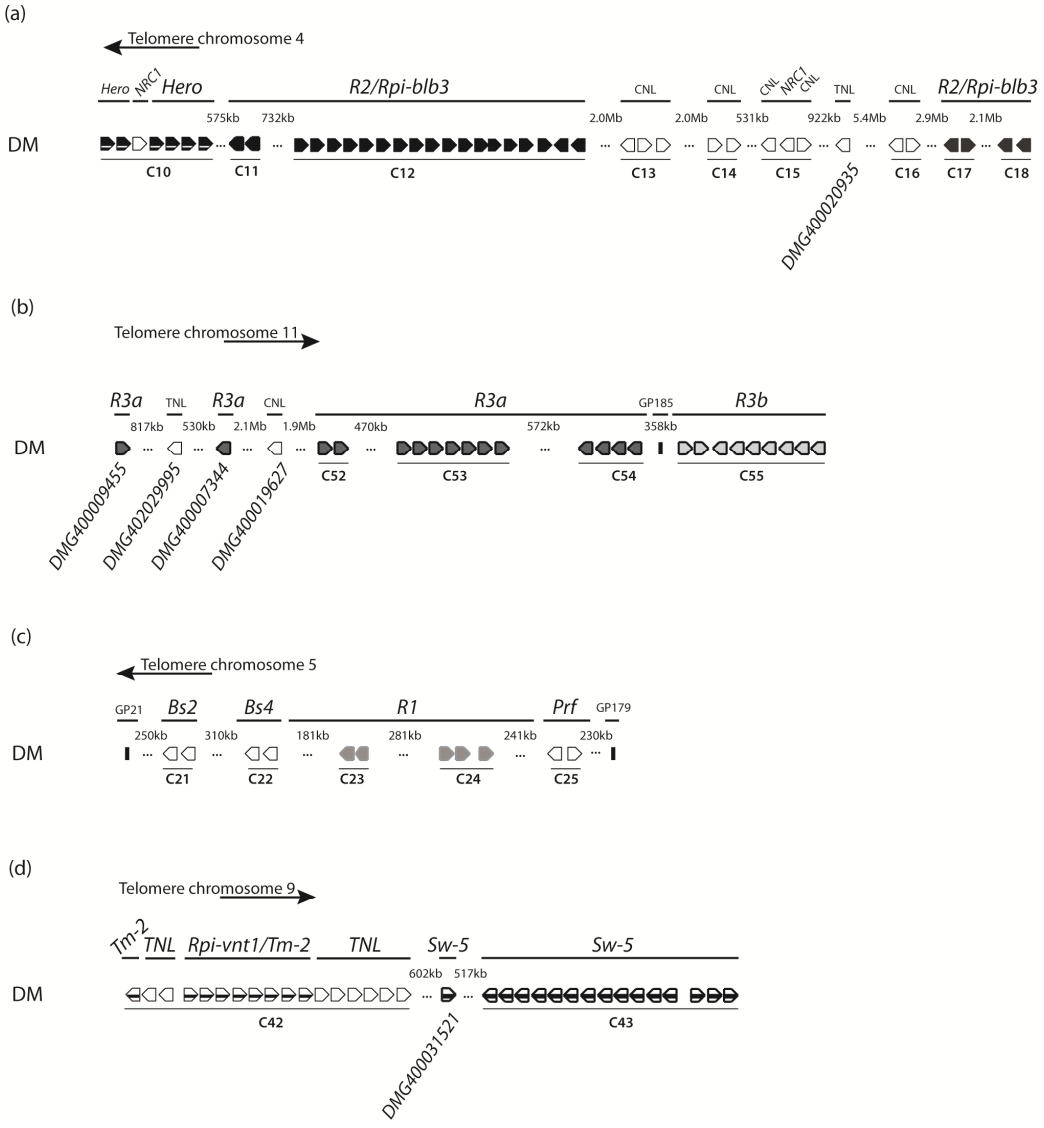


Figure 9 Physical overview of selected resistance gene loci: *R2* (a), *R3* (b), *R1* (c) and *Rpi-vnt1/Tm-2/Sw-5* (d). The directions towards the respective telomeres are shown. Boxed arrows symbolise NB-LRR genes and clusters are indicated by horizontal lines. Known genetic markers are shown. The distances between NB-LRR clusters are indicated above the gaps. Identifiers for single NB-LRRs are shown.

Previous studies have shown that the R1 resistance gene locus resides on chromosome 5 and is flanked by Bs4- and Prf-like R genes (Ballvora et al. 2002;Kuang et al. 2005). This structure has been maintained in DM. Four adjacent clusters (C22 – C25) contain two TNLs with homology to BS4 (C22), five R1 homologues in clusters 23 and 24, and two Prf homologues in cluster 25. Two BS2 homologues in cluster 21 (Figure 9c), are positioned approximately 310kb upstream of C22.

The long arm of chromosome 9 features two large heterogeneous clusters. Cluster 42 harbours eight TNLs that are separated by eight homologues of Rpi-vnt1 (Foster et al. 2009) and Tm-2 (Lanfermeijer et al. 2003). The more distal cluster C43 contains 15 homologues of the Tospovirus resistance gene Sw-5 (Brommonschenkel 2000) (Figure 9d).

2.2.5 Genomic organisation of NB-LRR genes

Gene and repeat densities were calculated and visualised for mapped gene features of the DM genome using a window size of 250kb centred on each gene in the corresponding superscaffolds. This analysis was carried out together with Dr Leighton Pritchard (JHI). DMGs for which the 250kb window would extend beyond a superscaffold were omitted from the analysis. Figure 10 indicates contours for a Gaussian mixture model (GMM) with two components that was fitted to the gene/repeat density data. The bulk gene/repeat density is modelled as two overlapping populations that are better distinguished in terms of gene density than repeat density. This is consistent with the potato genome analysis described by the PGSC (2011) indicating that there are relatively 'gene-rich' and 'gene-poor' regions within the DM genome. The GMM is overlaid in each case with a scatterplot showing data for predicted NB-LRR genes that were suitably placed for analysis within the superscaffolds. The majority of NB-LRRs lie within the contours of the GMM, consistent with the distribution of NB-LRRs being similar to that of all other genes in the potato genome. Only sixteen genes are visually distinguished as lying outside the contours of the GMM and mainly located in relatively repeat-rich regions. This number is within the statistical expectancy of sampling error. It is however interesting to note that eight of these genes are members of phylogenetic subgroup CNL-1: DMG 0025512 from cluster 27 and DMGs 0031878, 0020732, 0020735, 0020736, 0020740, 0020741, and 0020749, which are adjacent to one another in cluster 28. Phylogenetically, members of the subgroup CNL-1 are most similar to the *P. infestans* resistance gene *Rpi-blb2* and the nematode and aphid resistance gene *Mi-1* (Figure 6 and Table 4). Four further CNLs that are located in more repeat-rich regions are DMGs 0029453, 0029505 and 0029506, and all of them group together in the heterogeneous cluster C10 on chromosome 4 whereas DMG 0016372 is a single NB-LRR gene on chromosome 1.

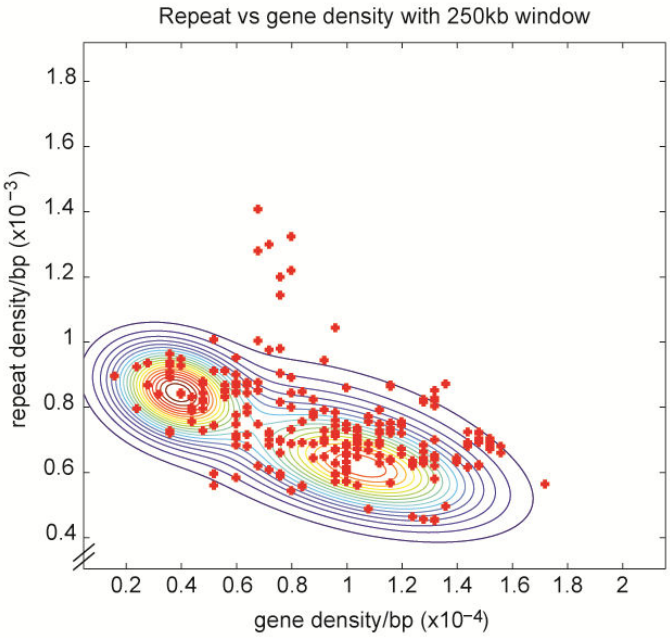


Figure 10 Global gene density versus repeat density analysis. The contours represent a genome-wide Gaussian mixture model (GMM) with two components fitted to the gene/repeat density data in a 250kb analysis window. Overlaid on the calculations are the CNLs and TNL type genes (shown as red crosses).

2.3 Discussion

We used an iterative process of manual and computational analysis to identify 438 NB-LRR-encoding sequences within the recently published doubled monoploid potato genome (PGSC 2011). This study has revealed a slightly higher number of CNLs and TNLs compared to the 408 NB-LRRs described by PGSC (2011), and 435 identified by (Lozano et al. 2012). The difference to PGSC, which is within the expected sampling error, includes 28 additional TNL genes and 2 additional CNLs. By extending the DM gene models by 3kb at the 3' and 5' end respectively to produce the DMG+ sequences, more domains associated with NB-LRR type genes were identified and the gene annotations correspondingly extended. The number of annotated partial NB-ARC only genes fell in our predictions from 105 to 41 (Table 3). Whilst our analysis used NB-LRR discriminative MEME motifs derived from a training set harbouring functionally characterised NB-LRRs from the wider plant kingdom, the analysis described by PGSC (2011) is based on NB-derived Pfam domain searches, followed by the construction of a potato-specific NB hidden Markov model. Both approaches yielded very similar numbers of NB-LRRs. Unfortunately, a direct comparison between the different resistance gene homologues (RGHs) was not possible as the identities of the CNL and TNL genes predicted by PGSC (2011) were not made publicly available. Analysis carried out by Lozano et al. (2012)

involved a similar approach as used by the PGSC, but ID's were made available for all identified sequences. These showed that the actual number of individual gene models identified was 387, and that 47 alternative splice variants had been included. The analysis reported here has identified 71 DMGs which are not present in Lozano et al. (2012). Furthermore, 23 DMPs reported by Lozano et al. (2012) were below the E-value threshold applied in the MAST search reported here and were thus discarded from the NB-LRR gene candidate list.

The MEME motif and phylogenetic analysis revealed a distinction between CNLs and TNLs in the N-terminal region, and in the NB-ARC domain of these sequences. Seven of the 20 identified MEME motifs (Table 2) distinguished between these NB-LRR subclasses, or between the canonical and RPW8-type CNLs (Rairdan et al. 2008). Phylogenetic analysis, which was performed on the conserved NB-ARC domain, supported this distinction and was consistent with previous observations for other plant species (Collier et al. 2011;Guo et al. 2011;McHale et al. 2006;Meyers et al. 1999;Meyers et al. 2003;Mun et al. 2009;Rairdan et al. 2008;Yang et al. 2008b).

The DM potato genome harbours 4.7 times more CNL than TNL genes. A similar distribution was found for the NB-LRR genes of grapevine (3.8×), but the ratio is smaller in poplar (1.7×) (Yang et al. 2008b). In comparison, the NB-LRRs of the Brassicaceae *A. thaliana, A. lyrata* and *B. rapa* contain CNLs and TNLs in a 1:2 ratio (Meyers et al. 2003, Guo et al. 2011, Mun et al. 2009). The genome of the monocot rice contains only CNLs; all other grasses analysed so far contain no or only very few TNLs (Tarr and Alexander 2009;Yang et al. 2008a). Leister (2004) suggested that overrepresentation of TNL over CNL genes in the Brassicaceae *Arabidopsis* and rape seed could reflect the adaptation of the *R* gene set to the predominant pathogens. It can be speculated that the over-representation of CNLs in potato is a response to some of the most damaging pathogens such as *P. infestans*, which is typically controlled by CNLs. In line with this, it is interesting to note that 27% of the identified NB-LRR genes share high sequence similarity to functionally characterised *Rpi* genes.

The CNL branch forms two phylogenetic clades, containing the canonical CNLs and the CNL-R (CC_{RPW8}-type), as previously described (Rairdan et al. 2008, Collier et al. 2011). Within the canonical CNLs, eight major subgroups with high support and short branch length were identified, suggesting a recent common ancestor. Two-thirds (13 of the 21) of the functional CNL genes included in the tree are found in these subgroups. Only members of CNL-3 and CNL-7 (and some of the smaller subgroups) show no significant sequence similarity to a functionally characterised *R* gene thus far. Their role, which is hitherto unknown, could for example be to provide resistance to yet unknown

pathogens and/or to mediate non-host resistance responses (Schulze-Lefert and Panstruga 2011).

In the context of the potato genome, the percentage of all genes that are predicted to encode NB-LRRs is 1.16% while the sequence makes only 0.15%. This is in line with estimates for other plant species that range between 0.6–1.8% (Mun et al. 2009). The gene density around potato NB-LRR loci is approximately 100 genes per megabase. However, unlike RxLR effectors from *P. infestans* which often reside in gene sparse- and repeat-rich regions (Haas et al. 2009), a global analysis of the DM NB-LRR genes (Figure 10) shows that CNLs and TNLs reside in genomic regions that are not significantly different to the potato genome in general in terms of gene or repeat density.

Several approaches for the identification of NB-LRR clusters have been described, and we have utilised a combination of the analyses described by Yang et al. (2008b) and Meyers et al. (2003). The identified members and the overall number of predicted clusters were very similar for both types of analyses, suggesting that the identification of clusters by these methods is relatively robust. However, cluster prediction based on the distances between NB-LRRs does not take into account the variability of gene density in the potato genome (PGSC 2011). Similarly, the definition of a gene cluster solely based on the number of non-NB-LRR genes between CNLs and TNLs fails to take into account any physical distance. Predicted potato NB-LRR genes are unevenly distributed over the 12 chromosomes and cluster into groups of different sizes. This is in line with data for other plant species (Meyers et al. 2003;Mun et al. 2009;Yang et al. 2008b). Various mechanisms including recombination, gene conversion, duplication and selection are thought to contribute to the genome-wide diversity and distribution of NB-LRR gene loci (Baumgarten et al. 2003;Friedman and Baker 2007;Leister 2004;McDowell and Simon 2006). Equal intragenic crossing-over results in domain swaps between genes whereas unequal crossing-over influences the number of genes within a locus and potentially places genes into a new structural context. Tandem duplications, in which the copy is contiguous to the original gene, are typically associated with homogeneous clusters. Of the 63 clusters, 50 are homogeneous and thus likely a result of tandem duplications. Members of the subgroups CNL-1 to CNL-8 are often found on the same chromosome and, in some cases, within the same clusters, which is consistent with tandem duplication. In contrast, segmental and ectopic duplications, which involve the duplication of entire gene blocks or single/small groups of genes respectively, can position copies to unlinked sites including different chromosomes (Leister 2004). Both CNL and TNL distributions display evidence for events that placed homologous genes on to different chromosomes that could be a result of either segmental or ectopic

duplication. These events appear to be more common for TNLs which are more widely dispersed throughout the genome and not found in clusters as frequently as CNLs.

The sequencing of DM provides a snapshot of the potato genome organisation, and specifically the distribution of and relationships amongst NB-LRR genes on individual chromosomes. Although specific to DM, this analysis provides an important starting point to gain insight into the NB-LRR gene compositions of other members of the Solanaceae. Studies in *Arabidopsis* have shown, for example, that some *R* genes display high levels of polymorphism within and between populations (Guo et al. 2011). A more detailed analysis of the potato *R1* locus (Ballvora et al. 2002), for which three haplotypes from *S. demissum* have been described (Kuang et al. 2005), confirmed evidence of copy number variations consistent with tandem duplications. As previously described, the *R1* locus is flanked by sequences similar to the tomato *R* genes *Bs4* and *Prf* but the number of *R1*-homologues varies between one and 17 in *S. demissum* and five in DM (Figure 9c; Kuang et al. 2005). Another example is the *R3* locus on chromosome 11 which was originally described in a diploid potato population, SHxRH (Huang et al. 2004). Overall, *R3* cluster organisation is syntenic between SH-haplotypes and the sequenced DM, in that the *R3a*-clusters (C52, C53, and C54 proximal) and the *R3b* cluster (C55, distal) flank the marker GP185 (Figure 9b). However, in DM, the physical distance between the clusters C54 and C55 amounts to more than 350kb and is thus approximately 200kb shorter than the same region in SH (Li et al. 2011). In DM, nine *R3b* homologues reside in cluster C55, whereas Li et al. (2011) describe six and ten homologues for the two SH haplotypes. Unequal representation of lineages within the NB-LRR superfamily and copy number variation between haplotypes is consistent with a 'birth and death' model in which some NB-LRRs are lost and new lineages evolve whilst others are retained (Michelmore and Meyers 1998).

The 438 NB-LRR genes described here were identified in a doubled monoploid potato, which represents a single haplotype. Potato cultivars and breeding lines are often heterozygous tetraploids, which exhibit tetrasomic inheritance during crossing. The high levels of structural diversity observed in homologous *R* gene clusters from different potato haplotypes (e.g. Bakker et al. 2011;Ballvora et al. 2002;Kuang et al. 2005), and the extremely high levels of sequence polymorphism observed in potato, imply that it is highly likely that any given tetraploid potato clone may contain as many as 1,600 distinct NB-LRRs in its genome. A key objective for future resistance breeding is to understand the allelic diversity of NB-LRR genes in potato. Such an objective will require application of high throughput sequencing technologies allied to advanced bioinformatic tools for assembling sequence data from very closely related genes.

2.4 Materials and Methods

2.4.1 Identification of NB-LRR genes

'Positive' NB-LRR and 'negative' non-NB-LRR sequence training sets were used with the MEME Suite psp-gen script (version 4.4.0) (Bailey et al. 2010) to encapsulate information about probable discriminative motifs in the positive set. Then, using the psp file as additional input, MEME was run on the positive training set to identify the 20 most significant motifs in the sequences (Table 2). A MAST search was then conducted on a combined dataset of all (~56k) predicted protein models (PGSC0003DMP.pep.v3.4) and the training sets (Figure 5). DMP sequences were considered to be candidate NB-LRRs if their reported MAST E-values were lower than the least E-value for any member of the negative training set. A manual inspection of DMPs with E-values above this threshold was conducted to identify potential false negative results. Sequences that contained at least two TIR/CC-derived motifs or three NB-ARC-specific motifs were selected for further analysis as described below.

DM gene models (DMG) corresponding to the identified NB-LRR like DMPs, were extracted from 'PGSC_DM_v3.4_gene.fasta'. DMG sequences were extended by 3kb at the 5' and 3' ends using the DM superscaffold sequences in 'PGSC0003DM.superscaffold.fa' to generate the DMG+ set of potato genes, which were translated into all six reading frames. The MAST search with the potentially discriminatory MEME models was repeated to identify potentially missing domains, and the DMG+ sequences manually curated to produce the DMP+ set of protein sequences. DM homologues to members of the positive Solanaceous training set were identified by a BLASTP (Altschul et al. 1990) search.

2.4.2 Mapping annotated DMGs and repeat densities to the pseudomolecules

All DM superscaffold locations were extracted from the spreadsheet PGSC_DM_ v3_2.1.9_pseudomolecule_AGP.xlsx, downloaded from the PGSC data sharing site at *http://potatogenomics.plantbiology.msu.edu/data* (accessed on 25-09-2011). All DMGs were mapped from the input file PGSC_DM_v3.4_gene.gff, and all repeat positions were mapped from the file PGSC0003DMB.repeatmasker.gff (both provided by the PGSC), to the pseudomolecules (Dr Leighton Pritchard).

Gene and repeat densities were calculated for each annotated gene, using a range of window sizes (50kb, 100kb, 175kb, 250kb, 350kb, 500kb) centred on that gene, and relative only to the superscaffold on which the gene were located. Only the parent

superscaffold was used because the 50kb spacer regions introduced into the pseudomolecules may not accurately represent the expected separation between superscaffolds. Gaussian mixture models were fitted to the observed frequencies of gene versus repeat density for all annotated genes, using 200 bins for each measure (Dr Leighton Pritchard).

Gene clusters were determined from PGSC_DM_v3.4_gene.gff when the calculated distance between NB-LRR candidates is less than 200kb (Yang et al. 2008b), and no more than eight annotated non-NB-LRR sequences are present between two consecutive NB-LRR sequences (Meyers et al. 2003).

2.4.3 Multiple alignment and phylogenetic tree estimation

The NB-ARC protein domain region was chosen for phylogenetic analysis as the multiple alignment was tractable. NB-ARC sequences that were not full length were manually checked for sequencing and assembly errors. After this screening step, sequences of less than 50% of the full-length NB-ARC domain were excluded. The multiple alignment was built from 466 re-annotated DMGs, including 33 annotated *R* gene sequences (Annex 1) using the Pfam (Finn et al. 2010) NB-ARC domain (Pfam entry PF00931) seed alignment (12 sequences) and associated hidden Markov model using the hmmalign program from the HMMER 3.0 package (Eddy 2008). Model selection, using the joint estimation of amino acid substitution model and phylogenetic tree topology, was carried out using the TOPALi package (Milne et al. 2009), resulting in the selection of a WAG+I+G model. This model was used to estimate a Maximum Likelihood phylogenetic tree using the PhyML package (Hordijk and Gascuel 2005). Bootstrap support was based on 100 bootstrap replicates.

Chapter 3 – NB-LRR sequence capture from the sequenced clone DM – proof of concept analysis

3.1 Introduction

NB-LRR encoding genes are key players in plant disease resistance and their presence, absence or allelic variation is decisive for functionality. Bulk segregant analysis is a powerful tool to more effectively map underlying resistances and is based on the arbitrary distribution of non-functional genes in the bulk resistant and susceptible plants whereas the functional gene(s) is only represented in individuals that make up the bulk resistant plants (Michelmore et al. 1991). The NB-LRR gene complement derived from the potato genome sequence has been detailed in chapter 2 (Jupe et al. 2012). This information forms a blueprint to develop a novel strategy to directly sequence and compare all members of this family from segregating bulks from wild or cultivated *Solanum* accessions. Sequence information from a single second-generation sequencing run can cover the potato genome several times, but the NB-LRR gene family makes up for less than 0.2% of the total genome (Jupe et al. 2012). Recent advances in sequencing related technologies include on-array or in-solution sequence capture using RNA-baits designed on the genomic targets of interest (Gnirke et al. 2009;Mamanova et al. 2010). Targeted sequencing of NB-LRR encoding DNA fragments has great potential to drastically increase the output of NB-LRR specific sequence information, while decreasing 'non-informative' content. The resulting higher sequence coverage facilitates calling sequence polymorphisms more accurately, which is especially important for plants with higher ploidy levels (Parla et al. 2011;Saintenac et al. 2011).

The aim of the work described in this chapter was to develop a NB-LRR capture assay which, in combination with second-generation sequencing, can be used to compare the NB-LRR complements of resistant and susceptible bulks. I describe how the information about the DM NB-LRR gene complement (described in Chapter 2) was used to design a RNA-bait capture library (hence, *bait-library*). Furthermore, successful capture of NB-LRR fragments from the sequenced potato clone DM did not only verify the DM NB-LRR complement, but also identified 338 previously unidentified and un-annotated NB-LRR encoding genes. This proof-of-concept analysis was only possible with the genome sequence of the DM potato clone available.

3.2 Results

3.2.1 Design of a target enrichment bait library

To capture NB-LRR encoding DNA fragments and thus to enrich libraries for NB-LRR genes prior to Illumina sequencing, a customized Agilent *SureSelect Target Enrichment System* (Agilent, US) was used. We designed a bait-library based on 523 NB-LRR-like sequences from one of the first available annotations version 3.0, of the sequenced potato genome (PGSC 2011) (this bait-library was designed prior to the NB-LRR identification described in Chapter 2). Further, functionally characterized NB-LRR sequences were included. From the predicted ~60,000 DM potato protein models (DMP) in v3.0, 498 putative NB-LRR gene sequences were identified by Dr Dan MacLean through Pfam-domains and further hmm-models in predicted coding sequences (This analysis was carried out prior to results obtained in Chapter 2).

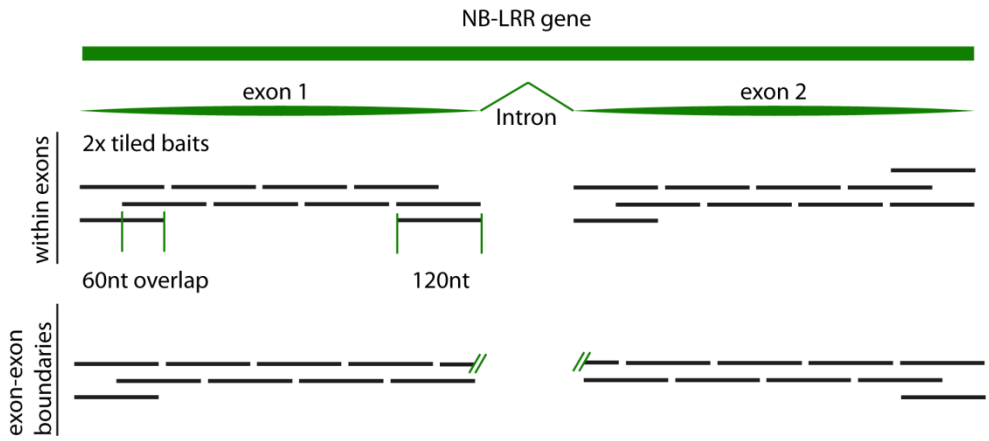


Figure 11 Single oligos for the NB-LRR enrichment bait-library are based on the previously identified NB-LRR like potato sequences. These 120-mers were designed to overlap 60nt into the previous bait and were designed within the single exons, but also over the exon-exon boundaries of the coding sequence.

Using an additional MEME and MAST search based approach (carried out by Drs Linda Milne and Leighton Pritchard; similar to chapter 2) 25 further novel sequences were identified and combined with the target gene library comprising 523 predicted coding sequences. The list was further expanded by incorporating nine, described in the literature, NB-LRR-type *R* genes from several Solanaceae, including tomato and tobacco. These were, *Bs2* (Tai et al. 1999), *Bs4* (Schornack et al. 2004), *Hero* (Ernst et al. 2002), *I2* (Ori et al. 1997), *Mi1* (Milligan et al. 1998), *N* (Zhang et al. 2009), *Sw-5* (Brommonschenkel et al. 2000), *Tm-1* (Ishibashi et al. 2007) and *Tm-2* (Lanfermeijer et

al. 2003). For tomato, a list of 57 NB-ARC domain containing sequences was generated from publicly available tomato sequences, and a total of 574 tomato derived oligos were added to the library. Computationally, two sets of 120-mer oligos were created, one set between and one set over the exon-intron boundaries, comprising 41,160 baits. After duplication of all NB-ARC specific sequences, 48,549 oligos were finally ordered as a customized RNA bait-library from Agilent. The oligos were designed with 60nt overlap, for each nucleotide to be represented twice, and with more overlap at the intron-exon boundaries or at the stop codons (Figure 11).

3.2.2 Analysis of Illumina sequence reads after DM NB-LRR capture

The customised Agilent SureSelect bait-library was used to capture NB-LRR sequences from genomic DNA of the potato clone DM. gDNA (Figure 12b) was sheared to produce nucleotide fragments between 100 and 1100bp in length, as can be seen in Figure 12c and after clean-up and size selection for fragments longer than 100bp using AMPure XP beads (1.8x volume) in Figure 12d. After repairing the ends and the addition of 3'- 'A'-overhangs, the DNA was cleaned using a lower concentration of AMPure XP beads (1.1x volume) resulting in the removal of fragments that are smaller than 180bp (Figure 12e). DNA fragments were subsequently ligated to paired-end Illumina adapters and, after a further clean up, used in a pre-hybridisation PCR to produce the correct Y-shaped Illumina adaptors (Bentley et al. 2008). The DNA was amplified with four cycles, after an initial PCR provided sufficient quantities (Figure 12g and h, respectively). The target capture hybridisation of the amplified DNA and the bait-library was carried out in half of the recommended volume for 36 hours. This is within the advised range of 24-72 hours. After capturing the RNA-DNA hybrids using Streptavidin coated beads and a digestive removal of the RNA oligos, a further PCR was carried out to amplify the captured library. As advised, a 10- and 11-cycle PCR amplification was compared (Figure 12i and j, respectively) and the amplification finally carried out for 10-cycles.

The enrichment efficiency was determined through comparison of the expression levels of *R2*- and *R3*-like genes by qPCR on DNA samples prior and after enrichment with NB-ARC domain specific primer pairs. A further primer pair was used which is able to amplify the NB-ARC domains of several homologues of a yet uncharacterised NB-LRR gene cluster (Walter Verweij, unpublished). The results (Figure 13) showed that for all primer combinations the enriched template entered the log phase around eight cycles earlier than the non-enriched (20 cycles and 30 cycles, respectively) after correcting for

the ΔC(t) of 2 in the endogenous control gene *18S*, suggesting a high enrichment of NB-LRR specific sequences.

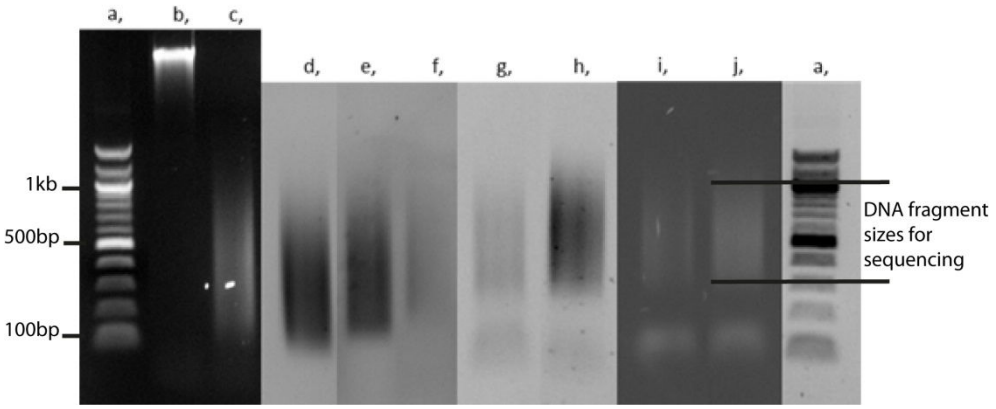


Figure 12 Agarose gel electrophoresis was used to observe DNA quantities and changes in fragment size during the Illumina paired-end sequencing library preparation and enrichment. In this figure are shown (a) 100bp ladder, (b) extracted DM gDNA, and (c) after shearing to a smear of 100-1000bp with a Covaris sonicator. (d) AMPure clean up post shearing, followed by (e) A-tailing and AMPure clean-up using higher concentration of AMPure beads removed fragments smaller than 180bp. (f) After ligation to Illumina PE-adapters and AMPure clean-up a further shift of fragment sizes can be observed. A comparison between (g) 4-cycle pre-hybridisation PCR and (h) 5-cycle pre-hybridisation PCR shows a difference in amplified template (equal amounts loaded). Post-enrichment (i) 10-cycle PCR and (j) 11-cycle amplification were compared and sufficient material for sequencing was produced with 10-cycles. (i) and (j) show the final fragment sizes that were Illumina sequenced, between 300 and 1000 bp.

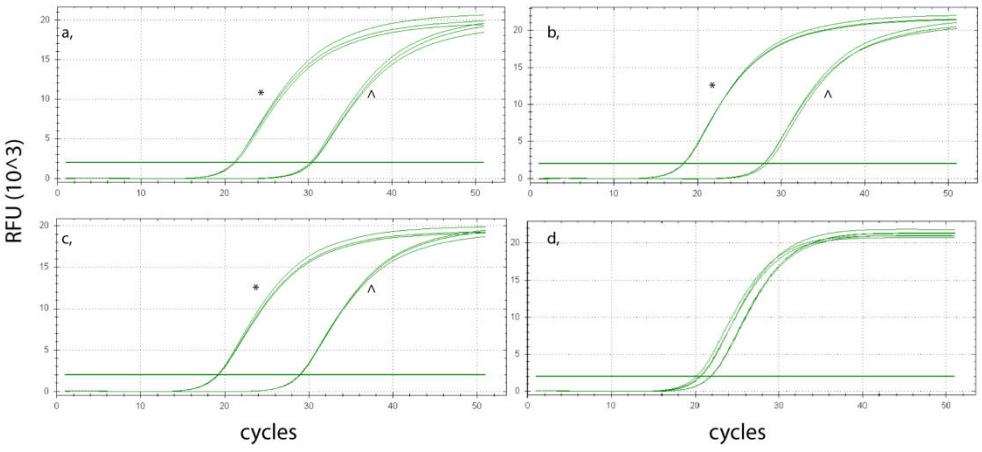


Figure 13 A qPCR was carried out to test the enrichment efficiency using primer pairs specific for the NB-ARC domains of (a) *R3,* (b) *R2,* (c) multiple NB-LRRs. As internal control, (d) *S. tuberosum 18S* was used. The amplification was carried out on gDNA sampled prior and post-hybridisation to the customized Agilent SureSelect bait-library. Horizontal green lines indicate the point when expression levels entered the log-phase, (*) marks the lines for enriched sample, and (^) for the non-enriched sample.

Paired-end sequencing of the captured library in a single lane of the Illumina Genome Analyzer II (Illumina, Inc.) generated 145.6 million raw 76bp long paired-end reads (11.06 Gb). The sequence data was subjected to stringent filter parameters using the Galaxy NGS TOOLBOX (Blankenberg et al. 2010b), in order to retain only high-quality reads. After removal of 5.6% of N-containing sequences and 46.9% that had an Illumina internal quality score below 20, 73 million high quality paired-end reads remained. Despite so many reads removed, the remaining sequence information would still, in theory, cover the whole DM genome 6.6-times.

The sequenced reads that passed the quality control were analysed through Bowtie mapping (Langmead 2010) to several references, allowing for two mismatches in the seed, and four in the extension: the DM superscaffolds, the pseudomolecules/chromosomes, the intended targets (predicted cDNA sequences used to design the bait-library), and the 438 DM NB-LRR sequences. Details are presented in Table 5. Of all reads, 88.9% could be mapped to the DM superscaffolds and 75.9% to the DM chromosomes. Those reads not mapping to a chromosome could be attributed to the unanchored superscaffolds. In total, 31.5% of all reads mapped to the genes the bait-library was designed on, and 29.8% were mapped to the 438 DM NB-LRRs. On average, DM NB-LRR genes are covered by 18,403 reads/kb. Paired-end insert sizes were determined to be between 77 and 1000bp, with a peak at 208bp. These results show that NB-LRR specific DM sequence fragments were enriched from the calculated genome frequency of 0.15% of all DNA encoding for NB-LRRs to 29.8%, using the customized bait-library.

Table 5 Results of paired-end mapping of enriched reads to the bait-library sequences, DM superscaffolds, 438 DM NB-LRR sequences and the chromosomes of the potato genome. Reads were mapped using Bowtie, allowing for two mismatches in the seed, and four in the extension. M=million reads.

Reference	# hits	%	average	min	max
Bait-library cds	23M	31.5	43587	82	221282
DM NB-LRR	22M	29.8	49689	4	235784
Superscaffolds	65M	88.9	33632	2	2635092
Chromosomes	55M	75.9	4618477	1971610	7366606

3.2.3 Low coverage of potential off-target genes

As discussed in chapter 2, NB-LRR proteins consist of several domains that can also be found in a range of other plant proteins including NBS-kinases or RLKs. To assess the extent of sequence reads that cover these potential off-targets, all reads were mapped to the sequences of the negative training set (RLPs, RLKs, etc.; as defined in Chapter 2). The results are presented in Table 6 and show that nine of these ten sequences had a close to zero coverage with less than 67 reads per gene, none of which mapped to the DM NB-LRRs. The only exception is the *S. tuberosum* NBS-kinase protein Z2 to which 2759 reads mapped, of which almost 50% (1346 reads) also map to various NB-LRR sequences.

Table 6 Illumina reads were analysed for enrichment of potential off-targets. Bowtie mapping of all reads to the previously defined negative training set identified a very low read coverage for most of the sequences. Only the potato NBS-kinase protein Z2 gene showed higher read coverage. Reads were also cross-checked whether they mapped previously to a DM NB-LRR, and this was only the case for Z2.

# reads	map to NB-LRR	don't map to NB-LRR	negative training set (Chapter 2)
27	0	27	AB219939.1|Tomato Lehsp100 ClpB mRNA heat shock protein
20	0	20	AF053993.1| Tomato disease resistance protein Cf-5
17	0	17	AF082890.1|Potato cystathionine beta-lyase
14	0	14	AF272367.1|Tomato verticillium wilt disease resistance protein Ve1
2759	1346	1410	AF281282.1|Potato NBS-kinase protein Z2 gene
22	0	22	AM411448.1|Potato mRNA Ran GTPase-activating protein 2
67	0	67	AY112661.1|Tomato systemin receptor SR160
18	0	18	AY793347.1|Tomato *Cf-2.1*
7	0	7	DQ056434.1|Tomato xyloglucan-specific fungal endoglucanase inhibitor
7	0	7	DQ674708.1|Tomato ethylene-inducing xylanase
0	0	0	EF396238.1| *N. benthamiana* RAN GTPase-activating protein 1 (RanGAP1)
11	0	11	U15936.1|*L. pimpinellifolium Cf-9*
13	0	13	Y12640.1|Tomato *Cf-4A* gene

3.2.4 Discovery of further NB-LRR candidates from unmapped reads

Overall, 29.8% of the quality controlled enriched sequences could be assigned to members of the NB-LRR gene family. In a first attempt to determine the content of the remaining 70% of the reads, a *de novo* assembly was carried out using the software tool Velvet (Zerbino and Birney 2008) followed by alignment-based analyses using blastn of the resulting contiguous sequences.

Assembly of all non-NB-LRR mapping reads retrieved 49,853 contigs with a k-mer length of 57. That means that words with the length of 57 base-pairs were the starting points

for contig assembly and gave the best results for the Velvet assembler to create contigs from the sequence reads (Figure 14)(Velvet v1.0.0). Further assembly of these contigs using the programme Geneious (v5.5 created by Biomatters, available from http://www.geneious.com/) joined the 36,042 contigs to 5,279 new and larger contigs with a length between 109 and 29,562 nucleotides. The remaining 13,811 contigs assembled by Velvet remained unchanged.

Geneious derived contigs were initially screened using Blastn (Altschul et al. 1990) searches of the 40 longest contigs (5,035 – 29,562nt) against the NCBI nr-database, and this identified 38 hits for mitochondrial or chloroplast DNA, and, interestingly, also two with high sequence similarity to NB-LRR genes. Among 20 contigs of around 1kb length were eight sequences with high similarity to the NB-LRRs *SH10*, *Gro1*, and *Tm-2*. These results clearly suggested that there might be more NB-LRRs yet unidentified in the genome.

Carrying out a blastn search of the remaining 5,219 Geneious assembled contigs against the 438 DM NB-LRR cds, returned at least one positive hit for 2,197 of them (Figure 14). A comparative analysis between the superscaffold positions of these contigs and the DM NB-LRRs excluded 324 contigs as they are part of a previously described sequence.

In an attempt to create longer contigs and potentially re-assemble novel NB-LRR genes in full, the genomic positions of the remaining 1,873 contigs, positive for a NB-LRR, were used to extract 1kb additional sequence information to both sides from the superscaffolds. These extended sequences were used in a further Geneious assembly step and yielded 178 longer contigs, while 978 sequences were not assembled into novel contigs. A MAST search, as described in Jupe et al. (2012), identified 42 NB-LRR encoding candidate sequences within these contigs (Figure 14).

Repeating the analysis over the remaining 13,811 Velvet derived contigs identified 3,549 sequences with similarities to NB-LRR genes, of which 3,153 could be positioned on a DMB superscaffold. These positions were used to extend the sequences by 1 kb to each side prior a six-frame translation and MAST search. From these contigs, 34 NB-LRR encoding sequences were identified, resulting in a total of 76 potential new NB-LRR candidate sequences (Figure 14).

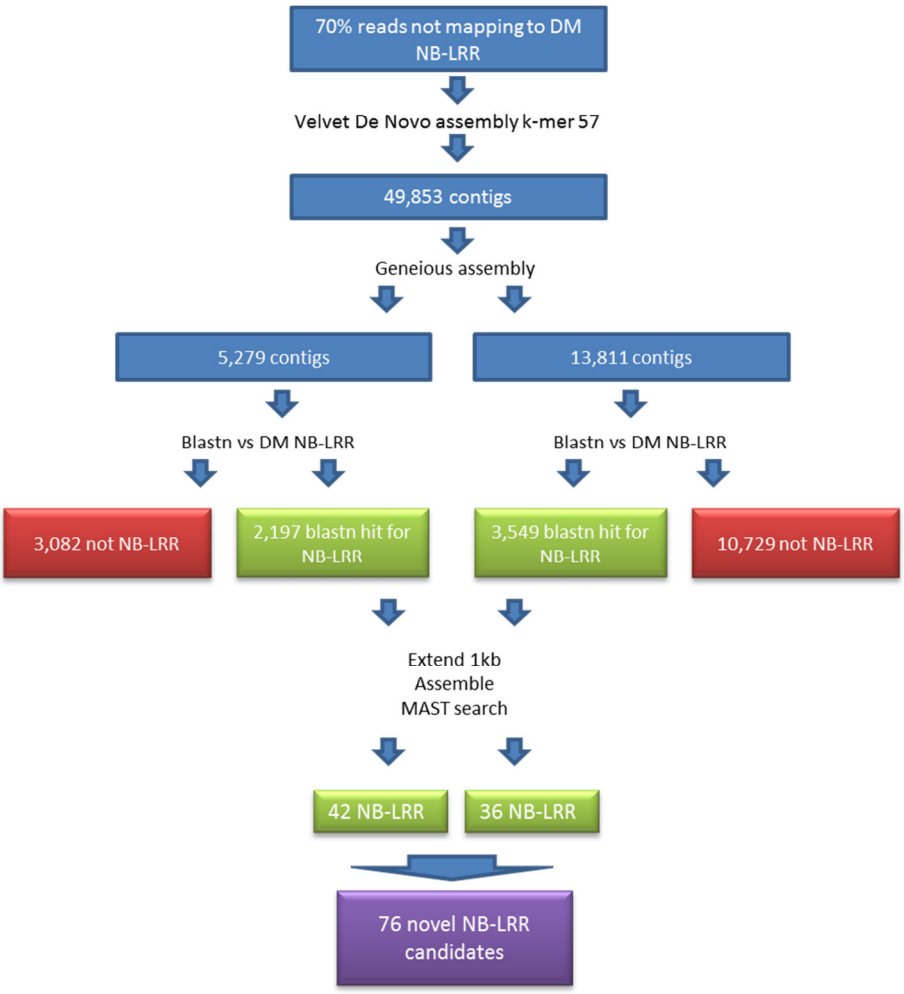


Figure 14 NB-LRR enriched DM Illumina reads that did not map to a DM NB-LRR gene were *de novo* assembled using Velvet, followed by a Geneious re-assembly. All derived contigs were used as queries in blastn searches against the 438 DM NB-LRR genes. Contigs with positive result were extended by 1kb based on positional information and again re-assembled. A MAST search over these contigs as described in Jupe et al. (2012) identified 76 new NB-LRR sequences that are positioned on the pseudomolecules and the unanchored superscaffolds of DM.

3.2.5 Regions of high read-depth unveil novel NB-LRR encoding sequences

The *de novo* assembly based analysis of post enrichment Illumina reads has identified a number of novel NB-LRR sequences. To further investigate what is encoded by those reads that could not be located to a NB-LRR, the mapping information of all Illumina sequence reads over the superscaffolds and chromosomes was visualised using the Geneious genome browser function. In addition, the positional information for the 438 DM NB-LRRs (chapter 2, Jupe et al. 2012) and all DM gene models (v3.4; PGSC 2011) was included. As the post enrichment sequencing reads were obtained from the same sequenced potato clone DM, very stringent mapping conditions were applied, allowing

for very limited mismatches (see Material and Methods). The initial observation provided further evidence that those regions not containing a NB-LRR gene have after the enrichment only a poor coverage compared to NB-LRR harbouring regions. Figure 16 shows an example on chromosome 11 for a region showing a NB-LRR (Figure 16b green bar) with high read coverage (a, log-scale and c, actual read alignments), followed by a region of continuous low read-depth where only non-NB-LRR genes were present (b, red bars).

A calculation of the read coverage for DM NB-LRR genes over the chromosomes identified a minimum of 25× over a 550nt window (Figure 15). For further analysis, all regions with more than 20× coverage over 500nt were extracted manually from the 12 chromosomes, or using a custom Perl-script from the unanchored superscaffolds (due to the larger dataset). The stringency of the conditions was adjusted to be more inclusive, and a total of 4,387 sequences were extracted from the chromosomes and 1,445 from the unanchored superscaffolds (Figure 15). Initially the chromosomal derived sequences were processed manually to remove those with an overlap to a DM NB-LRR, and pre-selected in a blastn search against DM NB-LRR sequences. This identified 367 new NB-LRR candidate sequences on which a MAST search was carried out, as described in chapter 2, and confirmed 211 sequences that harbour more than three consecutive NB-LRR specific motifs (Figure 15). This motif information was also used to determine spatial closely adjacent sequence fragments that have been extracted separately but belong to one gene. A further blastn search of the 156 contigs for which no significant MAST results were obtained showed to our surprise 61 false negatives that had clear sequence similarity to a known NB-LRR gene. Why these sequences have not been picked up by the MAST search remains elusive. Thus, a further 272 previously not identified NB-LRR genes have been located on the 12 DM chromosomes (Figure 15 and Table 7). Among these are 205 (75%) from regions without annotation or that are annotated as short/partial sequences by the PGSC, as comparisons with the DM gene model positions have shown.

Without an initial blast step, the 1445 sequences from the unanchored superscaffolds were directly analysed in a MAST search and identified 120 candidate sequences, of which 54 were identified to overlap with known DM NB-LRRs (Figure 15). Further seven sequences were at least twice as long as the annotated DMG and were subsequently re-annotated. This resulted in the identification of 66 new NB-LRR sequences from the unanchored superscaffolds (Figure 15 and Table 7).

Of the DM NB-LRR sequences predicted in Jupe et al. (2012) six (DMGs 0013095, 0029606, 0016628, 2002242, 0011112 and 0011115) were identified as duplicates of already annotated NB-LRRs. The previously predicted NB-LRR gene DMG 0033651 was discarded as it was not supported by any of our analyses and could no longer be found in the assembly. Further genes that could not be verified include DMGs 0007608, 0020913 (a and b), 0018046, 0018047, 0033159 and 2021043. However, in total, 424 previously identified genes could be verified, and together with the 338 newly identified NB-LRR genes, the revised and re-annotated DM NB-LRR gene complement comprises 762 sequences (Table 7 and Digital accompanying materials E3 and E6).

A close-up of a NB-LRR cluster on chromosome 11 (Figure 16d and e) shows an example region where high read depth aided the identification of novel gene models that encode for NB-LRRs. In this figure, green peaks (Figure 16d) indicate regions of high read-depth, while red bars below (Figure 16e) mark positions of previously identified DM NB-LRRs. Green bars (Figure 16e) below peaks harbour newly identified NB-LRR encoding sequences from previously un-annotated sequences.

All results were incorporated into a new chromosomal map of the DM NB-LRR gene distribution, based on the PGSC pseudomolecules version 3.2.1.10. With the recent change of pseudomolecule version (July 2012), positions were also established for v3_2.1.11, but not further used (see Digital accompanying material E3).

Table 7 New NB-LRR sequences were identified from regions of high Illumina read coverage following the NB-LRR gene enrichment, and are represented per chromosome or unanchored superscaffolds (DMB). A MAST and blastn search verified the sequence as candidate NB-LRR genes. Numbers in brackets indicate the number of previously identified DM NB-LRR genes that this analysis verified.

				candidates identified through			
	# DM reads mapping	Jupe et al.2012	extracted sequences	High coverage	MAST confirmed	blast confirmed	New NB-LRR
Chr01	6,410,348	28	446	7	4	7	35
Chr02	3,525,508	16 (15)	276	8	8	8	23
Chr03	2,701,644	4	228	0	-	-	4
Chr04	8,232,220	57	497	72	44	52	109
Chr05	5,828,546	27	356	42	26	38	65
Chr06	6,174,270	36	443	64	33	52	88
Chr07	3,481,762	13	274	5	4	5	18
Chr08	4,479,490	33(32)	283	17	12	15	47
Chr09	6,261,258	44(42)	419	28	20	21	63
Chr10	5,320,790	25(23)	377	39	12	15	38
Chr11	7,182,926	54	396	61	34	44	98
Chr12	5,285,258	33(32)	392	24	14	15	47
Unanchored	12,314,400	68(61)	1445	66	66	66	127
Total		**438 (424)**				**338**	**762**

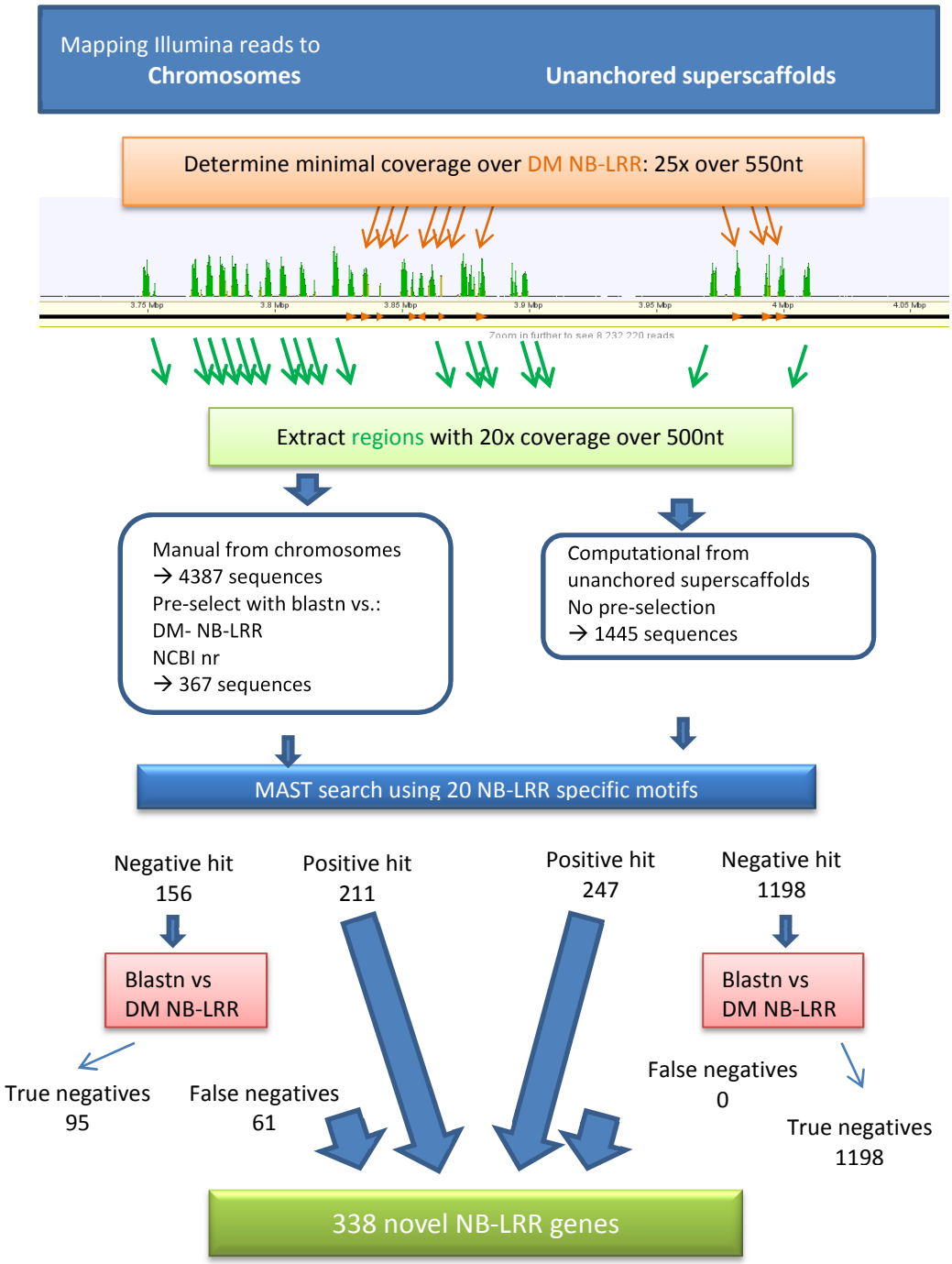


Figure 15 The NB-LRR enriched Illumina sequence read coverage, as well as positions of the 438 DM NB-LRRs were used to determine un-annotated regions with NB-LRR typical coverage (red arrows). Regions were subsequently extracted and analysed using blastn searches against the NCBI nr-database as well as DM NB-LRR sequences. MAST searches as well as further blastn of negative hits aided the identification of 338 novel NB-LRR loci.

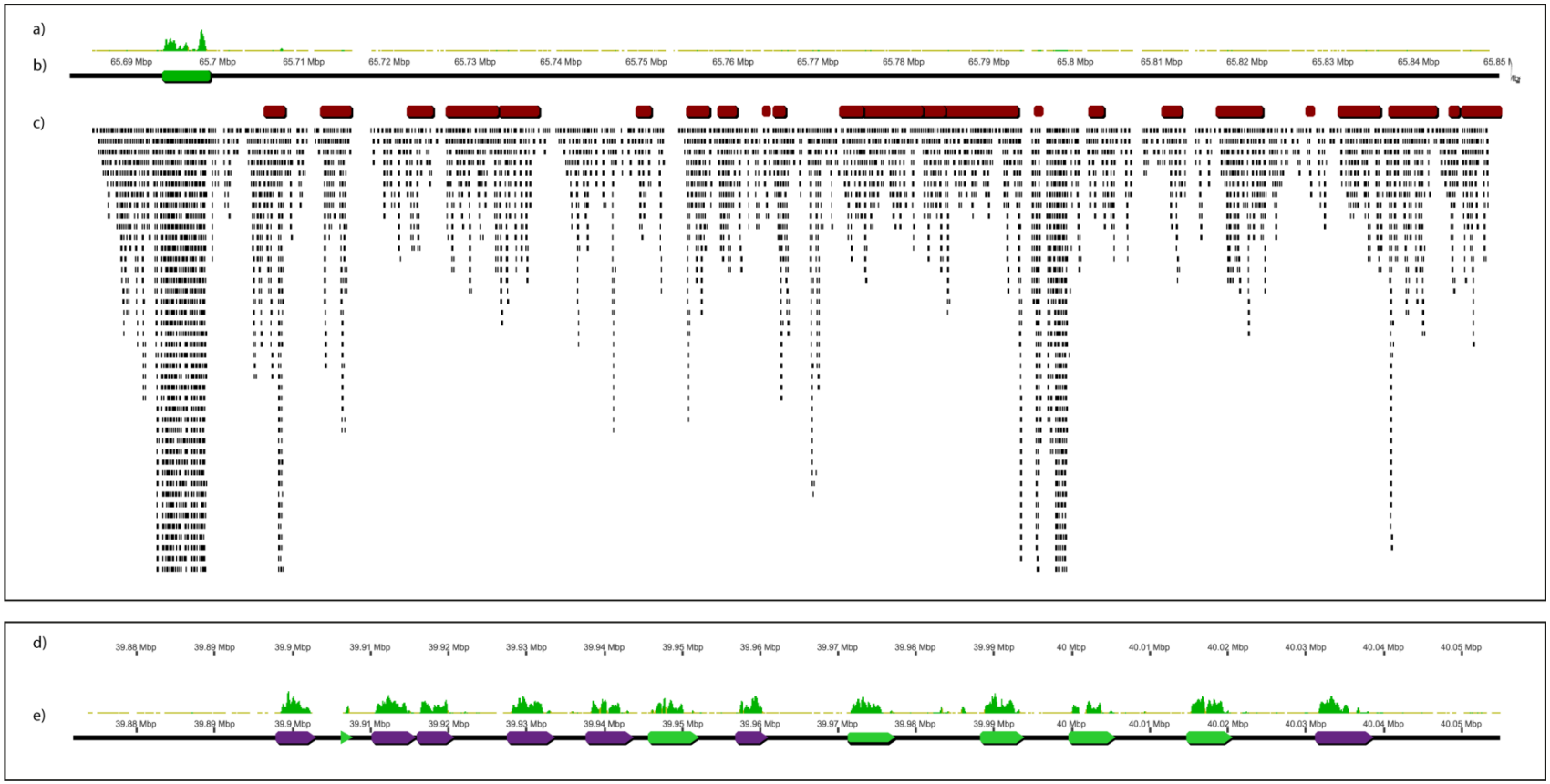


Figure 16 Close up of two NB-LRR clusters showing the density of background read coverage at the border (a-c) of the NRG1 region on chromosome 2, and an overview over the positions of some novel NB-LRRs within an annotated cluster on chromosome 11 (d-e). High read coverage is plotted in (a) as green peaks over a new identified NB-LRR, followed by non-NB-LRR DMGs in this area as red bars (b). Below (c) is the read depth plotted as single black lines depicting paired-end mapped reads. (d) Close up of cluster C77 of the R3 region between 39.9Mb and 40.0Mb on chromosome 11. Read coverage is plotted (d) as green peaks, while (e) red bars identify DM NB-LRR genes, and green bars newly identified NB-LRR sequences based on the high read depth above these regions.

3.2.6 Annotation of the expanded DM NB-LRR complement

The above analyses identified novel genomic regions harbouring newly identified NB-LRR sequences in previously poorly or incorrectly annotated regions by the PGSC. However, the analyses did not aim to establish detailed re-annotation as carried out in chapter 2 as this was beyond the time scale of this thesis. The finally established NB-LRR gene complement comprises over 2.06Mb of total sequence information if we assume the earlier established average length of 2.7kb per gene (chapter 2). This corresponds to 0.24% of the total potato genome. A basic analysis to identify the potential types of NB-LRRs was carried out by utilising a best-hit blastn search against the 438 DM NB-LRRs. This analysis showed that the CNLs are still the largest group of NB-LRRs and comprise approximately 584 members. The TNL subfamily has expanded to 157 members, whereas 21 new genes remain undetermined. The physical map positions have been updated where possible and contain now 635 NB-LRR gene positions over the 12 chromosomes (Figure 17). However, 17% of all NB-LRR genes are located on yet unanchored superscaffolds. A re-evaluation of the gene clusters showed an increase to 85 clusters containing 81% of all predicted NB-LRR genes (Table 8). Most clusters can be found on chromosomes 4, 5 and 6, with 19, 11 and 10 clusters, respectively. A good example for cluster expansion is the *R3* region on chromosome 11, where within C77 (former C53) six new genes were located resulting in an increase from seven to thirteen members. Similar, C76 (former C52) has expanded from two NB-LRRs to nine.

Table 8 Update of the NB-LRR gene clusters analysis reveals both expansions of existing clusters and the emergence of new clusters over most chromosomes.

			clustered genes		biggest cluster	
chromosome	# genes	# cluster	#	%	# members	ID
chr01	35	5	20	57	7	C53
chr02	23	5	16	70	6	C10
chr03	4	0	0	0	-	-
chr04	109	19	99	91	27	C17
chr05	65	11	51	79	14	C34
chr06	88	10	82	93	23	C42
chr07	18	5	11	61	3	C53
chr08	47	5	40	85	14	C60
chr09	63	2	44	70	25	C61
chr10	38	8	30	79	10	C68
chr11	98	9	85	87	19	C71
chr12	47	6	37	79	16	C84
Total	635	85	515	81		

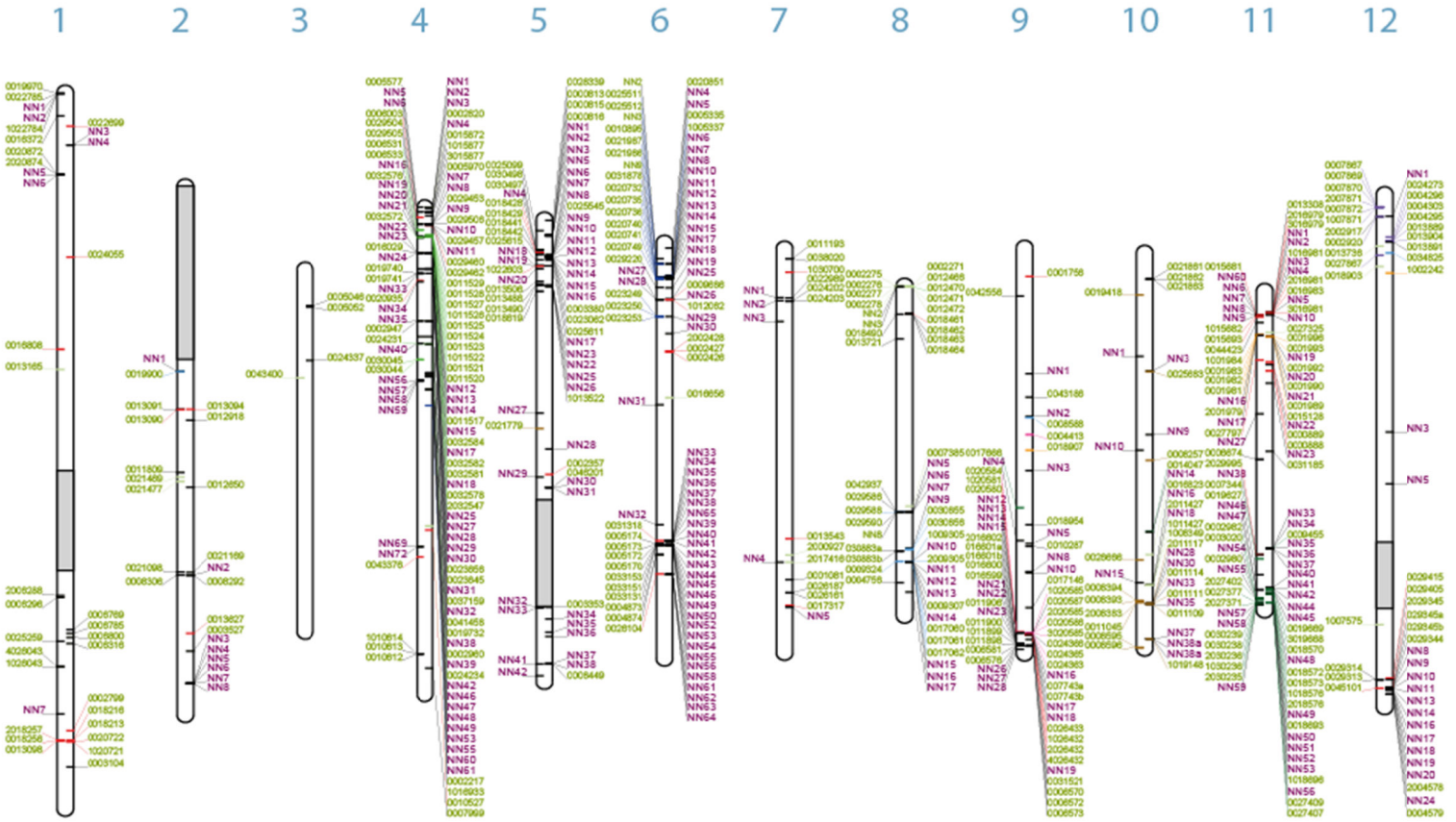


Figure 17 The 12 DM chromosomes are shown with all mapped NB-LRR genes. The newly identified genes are shown in purple, while all previously identified NB-LRRs are shown in green. Of all newly identified NB-LRR genes, 89% have been identified within known clusters or have been placed into new clusters.

3.2.7 Identification of novel NB-LRRs reveals higher enrichment efficiency

As shown above, the DM NB-LRR gene family has expanded by 338 new sequences. To re-evaluate the enrichment efficiency, all quality controlled paired-end reads (72.8 million paired-end reads) were mapped to the new set of 762 NB-LRRs. A total of 32.5 million reads, and thus approximately 44.6%, map to NB-LRR genes. Compared to the 0.24% NB-LRR complement over the total genome, NB-LRR specific reads have been enriched by approximately 186-fold.

Figure 18 shows the distribution of reads per NB-LRR, with the number of reads in 1000 on the X-axis compared to the number of NB-LRRs with this value on the Y-axis. The average number of corresponding reads per NB-LRR is 42,600. However the distribution is relatively even between 1,000 and 70,000 reads. Only one NB-LRR had less than 1000 reads, the TIR-only gene DMG 0004521 (24 reads). Interestingly, 20 sequences had more than 90,000 reads per gene, of which twelve had between 100,000 and 155,000 reads. The most reads were found for the CNL DMG 0004561 which had 217,842 reads, and is positioned on the unanchored superscaffold PGSC0003DMB000001060. Although some of these genes are situated within large clusters of NB-LRRs with high sequence homology, it could not be established why these genes were clearly overrepresented whilst others had fewer reads

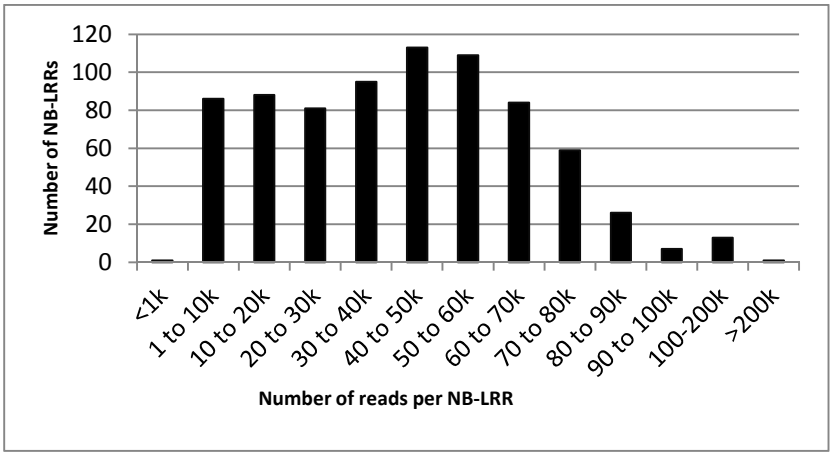


Figure 18 A total of 32.5 million reads map to the 762 NB-LRR genes of DM. The number of reads per NB-LRR was calculated and is plotted on the x-axis, with the number of NB-LRRs in the y-axis. k = 1,000

3.2.8 Enrichment efficiency in light of the bait-library design

The efficiency of the capture process is strongly dependent on the conditions and bait specificity during hybridisation. In theory, at a stringency allowing for null mismatches during hybridisation of the 120nt RNA baits with sheared genomic DNA, no NB-LRRs that are not present in the bait-library should be covered by a probe of the bait-library. *Vice versa,* under less stringent conditions and thus allowing for many mismatches, all somewhat similar genes should have a relatively high RNA bait coverage. An *in silico* experiment was carried out in which the approximately 49,000 baits were mapped to chromosome 11. An increase in the allowance for mismatches from 0% to 25% and 50% doubled the numbers of baits that could be mapped to chromosome 11 from 7,672 (16%), to 14,643 (30%) and 31,567 (65%), respectively (see Table 9).

Under very stringent conditions (0% mismatches), 23 (51%) of the 44 newly identified NB-LRR sequences on chromosome 11 had more than 5 corresponding baits. When the conditions were further relaxed to allow for 25% or 50% mismatches between the new NB-LRR gene sequence and the baits, this number increases to 31 (70%) and 41 (93%) sequences, respectively (Table 9). This is also depicted in Figure 19, where the change of bait coverage between old and new NB-LRRs is shown as a feature of the corresponding mapping stringencies. The fact that a number of baits could be aligned to a NB-LRR not used in the design of the bait-library might be due to the high sequence conservation within the NB-LRR subfamilies. Similarly, as the bait library was designed on an early genome annotation and did not encompass the 438 NB-LRR genes described in Jupe et al. (2012), the mapping analysis revealed that all previously identified DM NB-LRRs on chromosome 11 had at least one region with 6- and 8-fold coverage at 25% and 50% mismatches, respectively. As expected, a decrease in stringency leads to higher bait coverage of the NB-LRR genes on the chromosome.

An important conclusion from these analyses is that the combination of NB-LRR gene enrichment and sequencing in DM has identified additional NB-LRRs in previously poorly annotated genome regions. Furthermore, the enrichment procedure that enabled the discovery allowed for up to 50% mismatches between the 120 nt RNA baits and the hybridised target DNA fragments (Table 9).

Table 9 Summary of the *in silico* mapping of all bait-library sequences to chromosome 11 at three stringency steps. Shown is the number and percentage of baits that could be mapped, as well as the number of NB-LRRs they map to.

mismatch	# mapping baits	% of total baits	covered new NB-LRRs	not covered DM NB-LRRs
0%	7,672	16	23	4
25%	14,643	30	31	0
50%	31,567	65	41	0

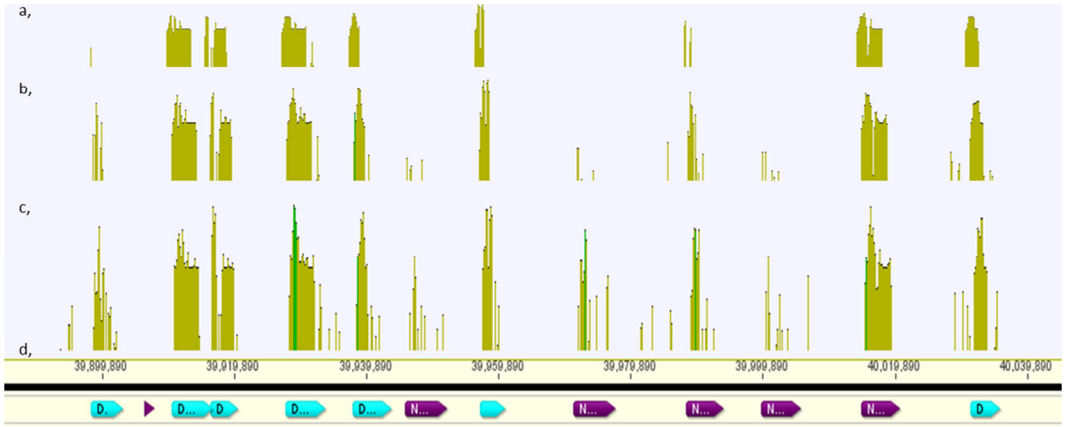


Figure 19 The bait-library was mapped *in silico* to DM chromosome 11 allowing for (a) 0, (b) 25 and (c) 50% mismatches. A visual comparison identifies large variations in the coverage (yellow peaks, green lines show high similarity matches) of previously identified DM NB-LRRs (blue arrows) and new NB-LRRs (purple arrows). All new NB-LRRs show mapping baits when at least 25% mismatch are allowed.

3.3 Discussion

An accurate and cost effective way to identify gene variations that are associated with diseases in humans is the targeted sequencing of specific genomic regions (Parla et al. 2011). A similar approach to target plant resistance genes has, to my knowledge, not yet been reported. NB-LRR proteins have been shown to be decisive for resistance towards a pathogen and thus alteration or absence of such a gene will lead to susceptibility. With the identification of the DM NB-LRR complement (Jupe et al. 2012) we created the basis to design and apply targeted sequencing of NB-LRR genes following bulk segregant analysis. This potentially aids the identification of variations in NB-LRR type resistance (*R*) genes of *Solanum* species.

With the release of the first DM protein predictions by the Potato Genome Sequencing Consortium in 2009, we were able to identify 523 NB-LRR encoding sequences. Because sequence capture efficiency is heavily dependent on the design of the bait-library (Parla et al. 2011), and some *R* genes in wild potato accessions might have more homology to *R* genes in tomato (Gebhardt and Valkonen 2001), a set of NB-ARC encoding genes derived from the first available versions of the tomato genome were also included, alongside nine characterised and diverse genes from various Solanaceae. Based on these sequences, a bait-library comprising around 49,000 120-mer RNA-derived probes was designed as a customized Agilent SureSelect target capture kit.

Before the first application of the enrichment kit on *Solanum* species or a population segregating for resistance to a pathogen, the efficiency of the NB-LRR capture was assessed on the reference potato clone DM itself. Although around half of the Illumina sequencing reads were removed during very stringent quality control steps, the remaining high-quality sequence information covered the potato genome theoretically 6.6-times. Mapping of the quality controlled captured sequences on to the targets used for the design and the later identified 438 NB-LRR genes showed coverage of 31.5% and 29.8% of the paired-end reads, respectively. The initially achieved enrichment efficacy is, however, far below for what has been shown feasible for Agilent human exome kits where ~70% of reads mapped to the regions used in the probe design (Parla et al. 2011). However, considering that the target region accounted for less than 0.2% of the total potato genome, the NB-LRR sequence enrichment was successful. The observed variation of target coverage depth is in accordance with reports by Parla et al. (2011) and seems to be an Agilent specific characteristic.

A first glance at contigs derived from a *de novo* assembly of the post enrichment Illumina reads showed evidence for traces of NB-LRR genes that were not within the mapping reference. The various NB-LRR explorations from the DM genome described so

far (Jupe et al. 2012; Lozano et al. 2012) were solely based on the gene and protein predictions of the Potato Genome Sequencing Consortium (PGSC 2011). The power of the enrichment procedure became apparent when the stringent mapping of the post enrichment DM reads yielded chromosomal regions within the DM genome that had previously not been annotated but showed significant read coverage. Indeed, upon further analysis of these regions, 338 new NB-LRR encoding sequences have been identified from mainly un-annotated regions of the genome. The number of post enrichment reads that could be mapped to the newly identified NB-LRR gene complement of 762 in DM increased to 44%. Based on the assumption that the 762 genes make up 0.24% of the DM genome, the actual enrichment efficacy is 186-fold. Together with an *in silico* mapping of the baits at various stringency levels, it could be determined that hybridisation during the NB-LRR capture allows for as much as 50% mismatches. Similarly, some baits did map under stringent conditions to NB-LRRs that were not included in the bait design which indicates that some baits are designed over conserved regions of the NB-LRR subfamilies. For example the *R* gene *Rpi-blb2*, which is a relatively recent gene and only found in *S. bulbocastanum*, for which we have identified 8, 1902 and 4156 baits from DM that would have covered this gene at 0%, 25% and 50% mismatch, respectively.

In human, as well as in plant research, a substantial number of SNPs has become available due to captured fragments being longer than the bait and thus reaching outside the exonic region into introns and 5' and 3' UTRs (Gnirke et al. 2009;Saintenac et al. 2011;Smith et al. 2011). Our analysis has shown that intronic regions as well as 5'- and 3'- UTR are partially covered. In our analysis sequenced insert lengths are 217bp on average and hence longer than the baits.

This analysis aided the re-annotation of the DM NB-LRR gene complement. It is interesting to note that 75% of the newly identified genes reside in previously not- or only poorly-annotated regions of the potato genome. This highlights a common problem with computational genome annotations and it appears that the gene model prediction algorithms used by the PGSC (2011) had potentially problems with the prediction of these highly similar and thus repetitive NB-LRR genes. The newly identified NB-LRR genes reside mainly within previously established gene clusters or form new clusters. A thorough phylogenetic analysis as described in chapter 2 has not yet been carried out; the consistent clustering of many genes and the conducted Blast searches suggest that many newly identified NB-LRRs belong to previously characterised subfamilies (Jupe et al. 2012).

Two approaches were used to discover the new NB-LRR set, a *de novo* assembly and an analysis of read coverage. The *de novo* assembly identified only a limited set of new genes from short contigs, while longer contigs over NB-LRR encoding sequences could not be made. The main problem is the high sequence similarity, short reads and the number of large gene sub-families. Given that DM has a monoploid genome *de novo* assemblies will most likely be unequally more difficult for potato species with higher ploidy levels. Very stringent mapping of the reads to the DM genome and identification of regions with high read coverage successfully identified the new NB-LRR set. But it remains to be seen whether DM can be used as a reference to assemble reads from wild *Solanum* species. The problem of allelic variation would however persist unless longer read lengths could be obtained. Indeed third generation sequencing such as PacBio RS (Pacific Bioscience, Ltd) can potentially yield 3kb sequence per read though the error rate remains too high according to a comparative study by Quail et al. (2012). Nevertheless, the results detailed here have shown that the RNA-based baits remain effective in capturing target sequences even when the sequence complementarity between the two molecules is as low as 75% or even 50% in extreme cases. This highlights the great potential for the application of this NB-LRR capture design to yield informative sequences from wild *Solanum* species that contain more diverse NB-LRR gene sequences.

3.4 Materials and Methods

3.4.1 Design of a customized Agilent SureSelect bait-library

Capture of NB-LRR like sequences was carried out using a solution based customized Agilent SureSelect Target Enrichment kit. To be able to design the bait-library on potato NB-LRR genes, a first draft of this gene complement was established using two different strategies in 2009. The first strategy is similar to the one described in chapter 2 and was carried out on version 3.0 of the genome annotation (Drs Linda Milne and Leighton Pritchard). The second strategy identified putative NB-LRR gene sequences through Pfam-domain searches and further hmm-models in the predicted coding-sequences of the same genome annotation (Dr Dan MacLean). For tomato, a list of NB-ARC domain containing sequences was generated using the previously applied hmm-model and a blastx search was conducted against the nr-database of NCBI to confirm sequence similarity to NB-LRR type *R* gene homologs (Dr Dan MacLean).

The bait-oligos were designed over and between the exon-exon boundaries of the motif containing sequences. The first bait started at the left most nucleotide and followed the predicted coding direction of the gene. Target sequences were at least two times tiled with 60nt overlap, and bait sequences containing Ns were removed. In order to utilize the full capacity of around 55,000 baits, oligos derived from NB-ARC domains of the potato genome were duplicated. In addition, nine Solanaceae NB-LRR type *R* genes were included. These were, *Bs2* (NCBI-accession AF202179) (Tai et al. 1999), *Bs4* AY438027 (Schornack et al. 2004), *Hero* AJ457051 (Ernst et al. 2002), *I2* AF118127 (Ori et al. 1997), *Mi1* AF039682 (Milligan et al. 1998), *N* EF091690 (Zhang et al. 2009), *Sw-5* (Brommonschenkel et al. 2000), *Tm-1* NM_001247615 (Ishibashi et al. 2007) and *Tm-2* EF137705 (Lanfermeijer et al. 2003). In total 48,549 120-mer RNA baits were designed to target a set of Solanaceae NB-LRR type genes.

3.4.2 Plant material, target capture and Illumina sequencing

DNA was extracted from leaves of a glasshouse grown *Solanum tuberosum* group Phureja clone DM1-3 516 R44 (DM) (kindly provided by Dr Sanjeev K. Sharma), using the Qiagen DNeasy kit according to manufacturer's protocol and quantified on a Nanodrop 1000 (Thermo Scientific).

A detailed protocol of the pre-hybridisation, hybridisation and post-hybridisation procedures can be found in Annex 2. Briefly, a total of 10µg gDNA was fragmented for 30s, 20% duty cycle, intensity level 5, and 200 cycles/burst (High-Performance Ultra-Sonicator, Covaris, Inc.). All subsequent steps were carried out according to the protocol of Agilent, with minor modifications (see Annex 2; carried out with help of Dr Walter Verweij). Sheared gDNA was purified using 1.8× AMPure XP beads (BECKMAN COULTER) and further processed for Illumina paired-end sequencing using the NEBNext DNA Sample Prep Reagent Set 1 according to the manufacturer's protocol (NEW ENGLAND BioLabs Inc.). After each step the fragment size was checked by electrophoresis on a 1.5% agarose gel and all fragments were purified using 1.1× AMPure XP beads (BECKMAN COULTER). DNA fragments were end-repaired, A-overhangs added and Illumina paired-end sequencing adapters ligated to them. Size-selection of the fragments was omitted, as AMPure XP beads removed fragments smaller than 180nt. A pre-hybridisation PCR was tested at four and five cycles using Herculase II polymerase (Agilent Technologies) with 10µl DNA template in a 40µl reaction mix. Visualisation on an electrophoresis gel showed sufficient amplification after four cycles. Finally, the PCR

was carried out at four cycles for the remaining 30µl DNA. Amplified DNA was mixed together, and combined with the hybridization block-mix containing 5µl plant capture enhancer (Roche Diagnostics Ltd) and 0.6µl Hyb-Block #3 (Agilent Technologies). Hybridisation to the custom RNA-baits took place in one half of the recommended reaction volume (14ul) for 36h at 65°C in a thermal cycler. RNA-DNA hybrids were selected using Dynal magnetic beads (Invitrogen, Life Technologies Corporation), purified with 1.1× AMPure XP beads and eluted in 15µl Sigma water (SIGMA-ALDRICH Co. LLC.). Post-capture PCR (Herculase II polymerase) was tested at 10 and 11 cycles with 5µl template each in a 20µl reaction mix, and finally carried out with the remaining 5µl template at 10 cycles. All PCR samples were combined and, after a final purification with 1.8× AMPure XP beads, eluted in 15µl Sigma water and measured using Quant-iT PicoGreen (Invitrogen Ltd) according to the manufacturer's recommendations. The higher concentrations of AMPure XP beads (1.8×) during the first and last purification steps aided the removal of fragments smaller than 180nt.

Enrichment efficacy was assessed in a quantitative PCR SYBR Green assay (Sigma-Aldrich Co.) with 1ng starting material from enriched and non-enriched samples, 0.5µl of the corresponding NB-ARC specific primer (100nM final concentration) and 8µl water. Primer information is disclosed in Table 10.

Table 10 Primers used in a qPCR to assess enrichment efficiency.

Target	Sequence 5' – 3'
NB-ARC forward	ACGAATTCGTTGTTGGTAGAGACAAAGATG
NB-ARC reverse	ACGGATCCGCTCTTAGTTTCTGACATTTCAGG
R3a NB-ARC forward	ACGAATTCAGAGCAGTCTTGAAGGTTGGAGC
R3a NB-ARC reverse	ACGGATCCATCTCCTTTCCGATTGCCACAAGG
R2 NB-ARC forward	ACGAATTCCAGCAGAGTCATTATTACCACG
R2 NB-ARC reverse	ACGGATCCAAGTAGTCCGCTCAATACAACAATTGC
18S forward	AACTTAAAGGAATTGACGGAAGG
18S reverse	AAGTTTCCCCGTGTTGAGTC

3.4.3 Raw sequence data processing

All quality control, mapping and sequence assembly experiments of Illumina paired-end sequences were carried out using scripts embedded in the TSL customized Galaxy instance (Blankenberg et al. 2010a;Giardine et al. 2005;Goecks et al. 2010), if not noted otherwise. Raw Illumina paired end reads were converted into FASTQsanger-format and quality controlled after joining the read pairs. Joint-reads shorter than 152nt, containing 'N', and/or having an Illumina quality score below 20 (allowing for 35 bases outside this range) were removed. Reads were then split and mapped using Bowtie (Langmead 2010) against the required reference using standard parameters, but allowing for two mismatches in the seed and four in the extension. The various potato genome sequence references were retrieved between March and July 2012 from http://potatogenomics.plantbiology.msu.edu/index.html. The read coverage over off-target genes was assessed by stringently mapping the left-hand reads to the negative training set, as defined in chapter 2. The conditions allowed for two mismatches in the extension, while other settings were default.

Numbers of mapping reads were generated in all cases using SAM-Tools (Li et al. 2009), and exported to a Microsoft Excel spread sheet for detailed analysis.

Genomic data was visualised with the genome browser function of Geneious (Geneious Pro 5.6, Biomatters; available from http://www.geneious.com/). Therefore a "genome" file (e.g. superscaffold, chromosome) was imported and subsequently overlaid with feature information (gff-format) including DM NB-LRR gene positions which were generated manually using Microsoft Excel.

Blast searches were generally carried out using default parameters, either as stand-alone Blast 2.2.24, embedded in Geneious Pro 5.6, or as NCBI blast in Galaxy.

3.4.4 Analysis of sequences not mapping to DM NB-LRRs

In order to identify the sequence content of reads that could not be mapped to the DM NB-LRR complement, quality controlled Illumina reads were *de novo* assembled using Velvet v1.0 (Zerbino and Birney (2008); embedded in Galaxy). Prior to Velvet assembly, the Velvet optimiser was run to identify the optimal k-mer length of 57. Resulting contigs were further assembled by the Geneious *de novo* assembler (Geneious Pro 5.5), applying default parameters in both cases. The 40 largest contigs, as well as 20 with a size range of 1kb were selected and used in a local blastn search against the NCBI nr-

database, and hits were curated manually. Several NB-LRR encoding contigs were revealed, and subsequently the 5,279 Geneious assembled contigs and the 13,811 remaining Velvet assembled contigs were used in a similar blastn search with parameters set to return only one hit per query. Positions on pseudomolecules or unanchored superscaffolds were determined by a blastn search against the corresponding database, and based on that, sequences were extended by 1kb to each side, and subsequently extracted. A further assembly in Geneious was only carried out for the pseudomolecule derived contigs, and 67 of the new assemblies showed one or more mismatches. In these cases the polymorphic sequences were removed manually. Final contigs from both pseudomolecules and unanchored superscaffolds were six-frame translated using a Perl script, followed by a MAST search as described in chapter 2 (Dr Graham Etherington). NB-LRR candidates were chosen when they had at least three consecutive NB-LRR specific motifs in one open reading frame.

3.4.5 Sequence coverage based NB-LRR identification

In the second approach, all reads were mapped separately to the twelve pseudomolecules PGSC_DM_v3_2.1.10_pseudomolecules.fasta, and to PGSC_DM_v3_2.1.9_superscaffolds_unanchored_gtr_2.5k.fasta. The resulting SAM-files were filtered for mapping reads and converted into BAM-format using SAM-Tools. The single pseudomolecules from PGSC_DM_v3_2.1.10_pseudomolecules.fasta were further uploaded into Geneious and the BAM-files were assembled to these. Regions of high coverage were identified in Geneious using the Coverage-tool, and candidate regions over the 12 pseudomolecules were identified visually based on a minimum length of 200nt. A similar approach was applied on the mapping information over the unanchored superscaffolds. Here, initially the coverage of previously identified NB-LRRs was determined for PGSC0003DMB000000008. A Perl script determined for known NB-LRRs a minimum coverage of 25 over at least 550 consecutive nucleotides (Dr Graham Etherington). For a more inclusive analysis, the parameters were reduced to 500nt minimum length with a minimum coverage of 20, and applied onto the BAM-file of unanchored superscaffolds using a Perl script (Dr Graham Etherington).

All identified sequences were used in a blastn search against the 438 DM NB-LRR genes, and if no hit was retrieved, against the NCBI nr-database. Sequences with NB-LRR specific hits were extracted and used in a MAST search (utilising previously determined parameters). A blastn search against the pseudomolecule and unanchored

superscaffolds files determined the position of the candidate sequences and this was compared with those of the 438 DM NB-LRRs.

The previously established DM NB-LRR.gff file was additionally loaded onto the pseudomolecule assembly to identify the correct positions of genes. In several cases DM NB-LRR positions were corrected as a result of scaffold inversions, and annotations were checked for overlaps of single genes. NB-LRR gene clusters were annotated as described in chapter 2.

Positions on the new version of DM chromosomes PGSC_DM_v3_2.1.11_pseudomolecules.fasta were established by blastn search of all sequences against this dataset.

All reads were paired-end mapped (Bowtie) to the new NB-LRR set of 762 sequences, and SAM-Tools was applied to determine the number of mapping reads per reference sequence.

3.4.6 *In silico* sequence search using the bait library

Hybridisation is thought to happen gap-free, but allowing for mismatches and loose ends. Therefore three *in silico* mapping experiments were carried out with the 48,549 baits on chr11. Geneious assembly was accomplished with the “Map to Reference” tool to chr11.fasta (word length 10, index word length 5, no gaps allowed, and maximum mismatch 0, 25 or 50%). Due to the design of the bait-library a NB-LRR sequence would be covered at least four-fold overall and eight-fold in the NB-ARC domain, and thus all regions with at least 4× nucleotide coverage were selected for further analysis. If coverage between these regions was within 3 of the highest coverage, these regions were connected and also selected. Selected regions were extracted from chromosomes and compared to DM NB-LRR positions to identify overlap.

Chapter 4 – Characterising the late blight resistance of Sarpo Mira

4.1 Introduction

The potato cultivar Sarpo Mira is known for its enduring resistance towards a wide array of *Phytophthora infestans* genotypes present in the UK and mainland Europe, including the most abundant and aggressive multi-locus genotypes (MLG) 6_A1 and 13_A2 (Cooke et al. 2012). Sarpo Mira is therefore used as a Europe-wide standard in field trials for late blight resistance screens and is also recommended for organic production (Colon et al. 2005; www.eucablight.org). The resistance against some isolates is, however, not absolute, as a restricted development of blight can be observed mainly at the end of the season. However, yield remains generally unaffected (Rietman et al. 2012;Sarvari Trust 2012)). Sarpo Mira was bred by Sarpo Kft., Hungary, and is derived from a cross between 76 PO 12 14 268 and D187, however further information about its progenitors has been lost, and the precise genetic and taxonomic origins of the resistance remains elusive. This red-skinned cultivar has been nationally listed in the UK in 2002 and is rated as a high-yielding late maturing main crop variety with additional resistances to the common viruses A, X, PLRV and Y (Shaw and Johnson 2004).

Genetic mapping of the underlying resistance from Sarpo Mira could be a very valuable resource for developing novel potato cultivars with enhanced resistance to the late blight pathogen. In comparison to the mainly diploid wild species used to clone resistance genes, Sarpo Mira is tetraploid and has already been optimised for commercial tuber production. Thus Sarpo Mira offers a much faster and efficient route to introgress the resistances into new cultivars if compared to utilising wild species as linkage drag of undesirable traits is greatly reduced.

Recent studies have attempted to pin-point Sarpo Mira's resistances using various strategies. Orlowska et al. (2011) measured and compared transcript expression from both a susceptible potato cultivar and Sarpo Mira during infection with late blight. Several differentially regulated genes were identified, among them the heat shock protein HSP70, ABC transporter, and WRKY transcription factors, all of which have been characterised during resistance signalling elsewhere (Orlowska et al. 2012). Interestingly, no NB-LRR genes were found among these differentially expressed genes,

suggesting that although some light was shed on the resistance pathway, the underlying receptor(s) which are assumed to be encoded by NB-LRR(s) has not yet been identified.

Rietman et al. (2012) tried to dissect the resistance of Sarpo Mira using an effectoromics approach. From a set of more than 200 cloned RXLR-containing *P. infestans* effector candidate genes that were transiently expressed in leaves of Sarpo Mira, five were recognized and triggered a hypersensitive response: *Avr3a* KI, *Avr3b*, *Avr4*, *AvrSmira1*, and *AvrSmira2*. Recognition of the first two Avr genes is conferred by the CNL type *R* genes *R3a* and *R3b*, respectively. It is hypothesised that the remaining *Rpi* genes *R4* and the novel *Rpi-Smira1* and *Rpi-Smira2* belong to the same class of genes (Rietman et al. 2012). Indeed, unpublished data suggests that *AvrSmira2* is likely to be *Avr8* which is recognised by the yet unknown *Rpi* gene *R8* located on chromosome 9 (Jo et al. 2011; Dr Jack Vossen personal communication).

The aim of this chapter is to gain some understanding of the incompatible interaction between Sarpo Mira and three contemporary UK late blight isolates of the MLGs 6_A1 and 13_A2, in light of the findings by Rietman et al. (2012). Additionally, it was intended to determine the genetic architecture of factors underlying the resistance against 3928A using a set of Illumina GoldenGate SNP assays. Lastly, a novel type of bulk segregant analysis was carried out that utilised NB-LRR capture and Illumina sequencing of bulked DNA.

4.2 Results

4.2.1 Characterization of Sarpo Mira's resistance to contemporary *P. infestans* isolates

Sarpo Mira is resistant to a wide range of *P. infestans* isolates, including those from the currently highly abundant MLG 6_A1 and 13_A2 (White and Shaw 2009;Cooke et al. 2012). Three contemporary late blight isolates were selected to dissect the resistance response of Sarpo Mira: 3928A from MLG 13_A2 which is known to have a short latent period and an extended biotrophic phase combined with high aggressiveness (Cooke et al. 2012), 7454A and 7822B from MLG 6_A1 (J. Lynott, personal communication). The virulence spectrum on Black's differential *R* gene series is for 3928A *R1–R4, R6, R7, R10, R11*, for 7454A and 7822B it is *R1, R3, R4, R7, R10* and *R11* (Cooke et al. 2012). In addition, 7822B overcomes *R8* (J. Lynott, personal communication).

A detached leaf assay was carried out with the isolates detailed above. These assays are high-throughput infection experiments that are commonly used for resistance screening through drop inoculations of detached leaves with a *P. infestans* zoospore suspension (Vleeshouwers et al. 1999). In contrast to available data from field tests (www.eucablight.org; Dr Alison Lees personal communication), detached leaves of Sarpo Mira showed typical late blight infection symptoms which developed as rapidly as on leaves of the susceptible control Maris Piper (Figure 20). Further resistance phenotyping was therefore accomplished through spray-inoculations with the same *P. infestans* isolates in whole-plant assays (Colon et al. 2004). These tests confirmed the resistance of Sarpo Mira and susceptibility of Maris Piper towards 3928A, 7454A and 7822B. Typical symptoms that developed after infection of Sarpo Mira with any of the three isolates were small black lesions that increased in size. On older leaves especially, lesions became wet and leaflets turned chlorotic. Those lesions however rarely sporulated or covered a whole leaflet. In contrast, Maris Piper developed the typical wet lesions associated with a compatible interaction after 3 days post inoculation (dpi) and sporulation was observed after 6 dpi in combination with total tissue breakdown.

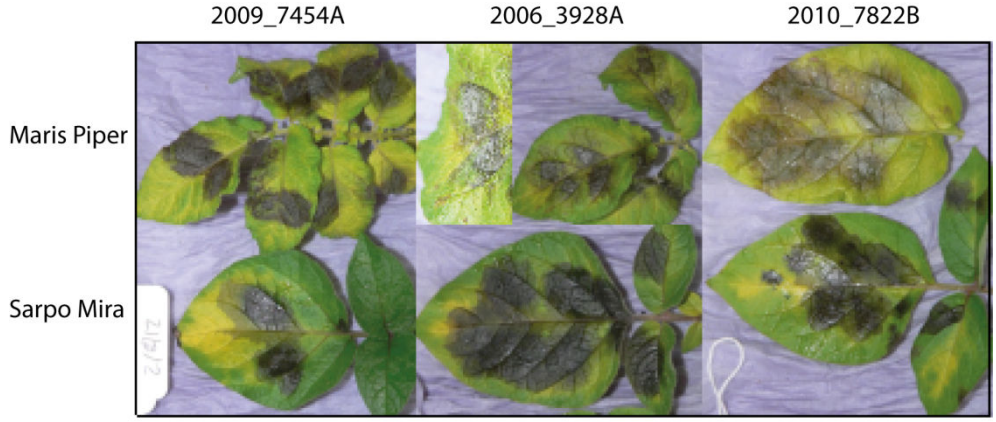


Figure 20 Detached leaf assay of Sarpo Mira and Maris Piper potato cultivars with *P. infestans* isolates 7454A, 3928A and 7822B. Photographs of leaves were taken 4 days post drop inoculation. Leaves were kept under high relative humidity. Sarpo Mira shows the same susceptibility level as Maris Piper.

4.2.2 Differential segregation of resistance towards 3928A and 7454A in a mapping population

The resistance response of Sarpo Mira to the two 6_A1 isolates was phenotypically identical, but different to 3928A. An existing F1 population derived from a cross

between Sarpo Mira and Maris Piper (SM×MP) was tested for differential segregation of the resistance towards 3928A and 7454A in independent duplicated whole plant assays (Colon et al. 2004). The isolate 7454A was used to exclude the involvement of *R8* in the resistance response.

In the screening for 3928A resistance (carried out in 2010 and 2011), 154 clones were spray inoculated alongside clones of the parental genotypes Sarpo Mira and Maris Piper. Sarpo Mira plants showed minor signs of a delayed lesion development 8 dpi as seen before, and were scored 8 on Malcolmson's 1 to 9 scale of increasing resistance (Cruickshank et al.1982), as shown in Figure 21a. Maris Piper plants showed high infection levels on the majority of leaves, and were scored on average as 3. Reproducible results were achieved for 129 clones, of which 17 were scored susceptible with scores between one and three, 14 were scored resistant with scores for both replicates of higher than seven. The remaining 98 clones were ranked as intermediates with scores between four and six (Figure 21a).

This bell-shaped and slightly skewed distribution (Figure 21a), with a mean of 5.2, indicates that the resistance of Sarpo Mira against isolate 3928A is not inherited as a simple trait, but has a more complex genetic basis. The above mentioned most resistant and susceptible plants were used for bulk segregant analysis which will be described in the following sections.

A further whole plant assay was carried out on 140 clones of the Sarpo Mira × Maris Piper population, using the blight isolate 7454A, and yielding reproducible results for 114 clones. Eight days post inoculation Sarpo Mira was scored 8 and 9, while Maris Piper was scored 2. In total 66 plants were scored resistant with scores of 7 to 9, while 34 plants had scores of 4 or lower (Figure 21b). Medium resistance, meaning clear symptoms of late blight infection on a smaller number of leaves was identified for 14 clones. This result is close to a 1:1 segregation of resistance, suggesting a single major *R* gene or resistance factor to be primarily responsible.

A summary of the results towards both isolates and overlapping resistance scores, based on the smaller bulks derived from the screen with 3928A, is shown in Table 11.

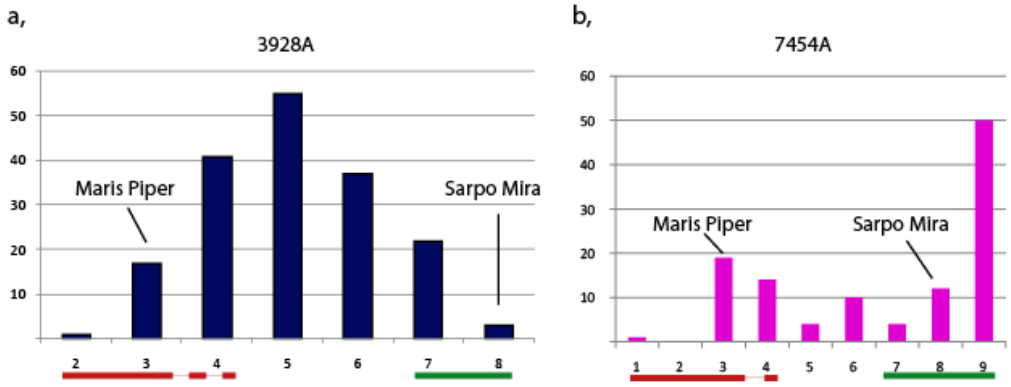


Figure 21 Average blight scores for a duplicated whole plant disease resistance assay involving the SM×MP population and infection with the *P. infestans* isolates a, 3928A or b, 7454A. Foliage blight scores taken 8 dpi on a 1-9 scale of increasing resistance are plotted in the X-Axis, over the number of plants in the Y-Axis. The graph also includes the scores for the parents. The red and a green bar indicate the most susceptible and resistant individuals, respectively, that were selected for a bulk-segregant analysis. Resistance scores of 4 (gapped red bar) were counted as susceptible, however not used for the bulks.

Table 11 The phenotypic scores of plants that were identified to be either resistant or susceptible to the two late blight isolates used in this study are set comparatively side-by-side.

	Response to	
SM×MP ID	3928A	7545A
194	S	S
21	S	S
90	R	S
7	S	S
165	R	S
196	S	S
88	S	S
52	S	R
14	R	R
123	R	R
61	S	R
98	S	R
187	S	R
94	R	R
100	R	R
104	R	R
141	R	R
160	R	R
180	R	R
197	S	R
221	R	R
SM	R	R
MP	S	S

4.2.3 Exploiting effector knowledge to determine resistance of Sarpo Mira

Rietman and colleagues attempted to dissect the resistance of Sarpo Mira through the use of transient expression of 234 predicted RXLR effectors (Rietman et al. 2012). From the recognition of five distinct effectors, they concluded the presence of the corresponding *R* genes in Sarpo Mira: *R3a*, *R3b*, *R4*, *Rpi-Smira1* and *Rpi-Smira2*. Two questions arose with respect to the underlying studies: a, are the corresponding effectors present in the tested *P. infestans* isolates, and b, does recognition of them co-segregate with the resistance response seen in the SM×MP population. In the following, an initial analysis on effector presence is presented. Preliminary data on effector recognition assays, carried out to answer the second question, were however not reproducible and are not presented in this thesis (see *Chapter 6 – Future perspectives*).

Virulence assays undertaken by James Lynott on plants of Black's differential *R* gene series and partially reported in Cooke et al. (2012) have shown that 3928A was unable to infect plants containing *R5* or *R8*, while the 6_A1 isolates are not virulent on plants harbouring *R2*, *R5* and *R6*, and in the case of 7822B *R8*. Several datasets on *P. infestans* gene expression derived from microarray experiments are available for various isolates, including T30-4 (Haas et al. 2009) and 3928A (Cooke et al. 2012). Cooke et al. (2012) deduced the expression of *AvrSmira1* and *Avr3a* at 2 and 3 dpi among 43 further RXLR effector genes. The expression coincides with the extended biotrophic growth phase of 3928A. Their analysis further showed that the present *Avr3a* homolog is homozygous for the amino acid substitutions EM, and thus cannot be recognized by R3a (only the KI-allele can be recognized by R3a; Armstrong et al. 2005). Notably, Rietman et al. (2012) reported that *AvrSmira2* was only expressed weakly, and at the very early time-point of 17 hpi. However no information is available for this time point for 3928A from the above mentioned studies.

To gain knowledge about presence and diversity of these effectors in the 6_A1 isolates used, full length genes were amplified from gDNA, cloned and sequenced. cDNA would have been the material of choice for expression analysis, however time constraints at the end of this thesis did only allow for the preparation of these experiments (see *Chapter 6 – Future perspectives*).

Sequence analysis of *Avr3a* showed that all three isolates are homozygous for the EM-allele and are thus not recognized by *R3a*. This analysis further identified the *Avr3a* paralogue *PEX147-3* in 3928A, which was previously reported to be recognized by *R3a.* Available microarray data presents the expression of *PEX147-3* at 3 dpi (Cooke et al.

2012). However, probes on the microarray might not have been specific enough and therefore cross-hybridisation might have taken place, as otherwise recognition should be observed in Sarpo Mira. Previous studies on other isolates have never shown expression for *PEX147-3*, although present in the genome (Dr Ingo Hein personal communication). Another paralogue, *PEX147-2*, was encoded by one of the sequences derived from 7454A, as well as a pseudogenized version lacking the stop codon in 7822B. Although *PEX147-2* is expressed (Cooke et al. 2012) for a further late blight isolate 88069 it was shown to be unstable *in planta* (Dr Ingo Hein personal communication). The effector *Avr3b* could not be amplified from any of the isolates, neither in full or using primers for a shorter amplicon, and the functional form is thus assumed to be absent. Sequence analysis of *Avr4* showed that all three isolates harbour the same putatively non-functional pseudogenized version with a premature stop-codon at amino acid residue E126. *AvrSmira1* could be amplified from gDNA of all three isolates, and sequencing revealed several synonymous and non-synonymous SNPs compared to the reference sequence of T30-4, as seen before by Rietman et al. (2012) who showed that all variants remained functional. The retrieved sequences for *AvrSmira2* are in all three isolates identical to the reference sequence from T30-4, which was tested by Rietman et al. (2012). Taken together, the presence of *AvrSmira1* and *AvrSmira2* makes them potential candidates for triggering the resistance seen against isolates 3928A, 7454A and 7822B. *Avr3a*, *Avr3b* and *Avr4* are not likely to be part of the response of Sarpo Mira seen in this study.

4.2.4 Genotyping the segregating population

Marker assisted selection exploits variations in genomic regions that are linked to a specific trait to screen breeding material rapidly through molecular analysis. To be able to use the resistance of Sarpo Mira against late blight isolates in marker assisted selection, markers need to be identified that are closely linked with the resistance response. Furthermore, the genetic position will enable the examination of the region for presence of a NB-LRR candidate, using the knowledge gained in chapters 2 and 3.

Identification of genetic linkages was facilitated by genotyping 179 clones of the SM×MP population alongside their parents. Genotypic information was gained for 1152 SNP markers derived from the sequenced DM potato genome and EST sequences from diverse potato cultivars (Dr Sanjeev K. Sharma personal communication) using the GoldenGate Technology (Illumina, Inc.). Groups of 384 markers are stored in five so

called ‘POPAs’ (*Potato Oligo Pooled Assays*), of which POPA 1, POPA 2 and POPA 5 were used in this study. The assay was out-sourced to the Sequencing and Microarray facility at The James Hutton Institute, which was provided with gDNA for the relevant samples. Briefly, in this highly multiplexed assay two locus-specific oligos for each of the 384 SNPs per POPA are hybridized to the DNA of each clone. Dye-labelled primers specifically bind to one of the two alleles and are further hybridized to a so called BeadChip. Colour read-outs from this BeadChip identify the allele as well as homo- or heterozygosity and give a specific theta score which combines the scores for the two alleles for each tested SNP. This data was exported in spread sheets for further analysis.

The derived theta scores were quality controlled by Dr Christine Hackett (Biomathematics and Statistics Scotland) using a multivariate analysis, and identified 13 samples with different patterns from POPA 1, and 23 from POPA 5. As outliers, these samples were unsuitable for further analyses.

The linkage groups were determined together with Dr Christine Hackett, for each parent separately based on segregation of single markers (Hackett et al. 2007). The identification of the chromosomal location for 1083 markers from the DM pseudomolecules v3_2.1.10 aided the consolidation of the identified linkage groups to the reference chromosomes. Further analysis determined 381 (35.2%) of the 1152 SNP markers, to be polymorphic in the population (Dr Christine Hackett)(see Table 12). The highest number of polymorphic markers was located on chromosome 4 and chromosome 1 with 50 and 42, respectively. Chromosomes 12 and 11 had the lowest number of polymorphic markers, with 14 and 15 respectively (Table 12).

Genetic linkage maps, shown in Figure 22, were created with the help of Dr Christine Hackett using JoinMap (Kyazma B.V.). Recombination frequencies and lod scores were derived from JoinMap for simplex markers in coupling phase, and chi-squared tests were used to link the double-simplex and duplex markers to these (Dr Christine Hackett). The coverage and density of the reference chromosomes with polymorphic markers was retrieved through alignments of the marker positions to the 12 DM chromosomes, and was found to be highly variable. For example chromosomes 1, 3 to 5, and 8 to 12 are covered over more than 91% of their length with polymorphic markers, although with much lower density on chromosome 12 (Figure 22b). While markers are spread over 85% of chromosome 7, this number is much lower for chromosomes 6 and 2 with 75% and 55%, respectively. No polymorphic markers were identified within the last 8Mb and 12Mb of chromosomes 6 and 7, respectively.

Table 12 Number of POPA SNP markers per DM chromosome as identified from DM_pseudomolecules_v3_2.1.10. After determining the individual linkage groups for each parent, polymorphic markers between them were identified, and are shown in the table.

Chr	POPA SNPs	Polymorphic (%)
1	134	42 (31.3)
2	120	40 (33.3)
3	90	31 (34.4)
4	134	50 (37.3)
5	95	27 (28.4)
6	72	25 (34.7)
7	81	34 (42.0)
8	77	34 (44.2)
9	79	34 (43.0)
10	81	35 (43.2)
11	63	17 (26.9)
12	57	14 (24.6)
Total	1083	383 (35.4)

4.2.5 Several QTLs are responsible for resistance towards 3928A

In order to identify markers that are linked with the complex resistance towards the late blight isolate 3928A, a regression analysis of resistance scores to the mapped markers was carried out by Dr Christine Hackett. While all effects with $p < 0.05$ are reported here, multiple testing effects have been ignored in this analysis. A detailed example of such an output is given for chromosome 2 (Table 13), while for all other chromosomes only the final results are presented.

In addition, the distribution of linkages that are specific to the most resistant plants used in the bulks was compared against all other values and tested using a chi-square test (Dr Christine Hackett). The results are presented in the following section for each chromosome, and shown in explanatory details for chromosome 3, where in order of the chromosomes the first statistically significant result was found.

For simplicity, markers are described in the text as numbers between square brackets, for example SNP [123]. The full identity of each informative marker is detailed in Table 15. Where dosages vary between the parents, Maris Piper is named first. For example duplex by simplex means that the marker is duplex in Maris Piper and simplex in Sarpo Mira. All results are summarised in Table 15.

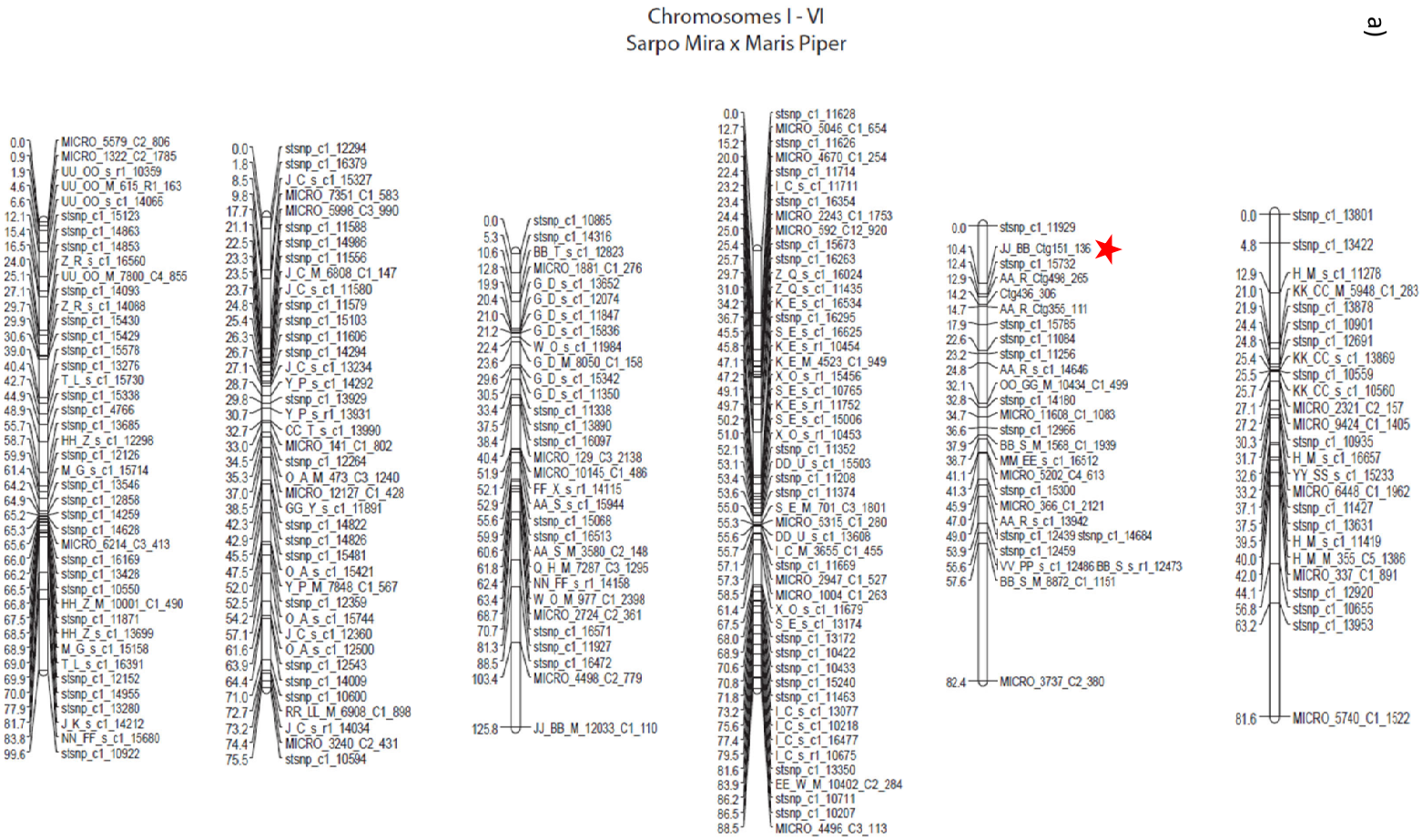


Figure 22 Genetic linkage maps for chromosomes 1 to 6 (a) and 7 to 12 (b). The marker position is given in centi Morgan to the left of each chromosome, and the name of the polymorphic POPA marker is given on the right. Markers that were later identified to be linked to resistance are marked with a red star.

b)

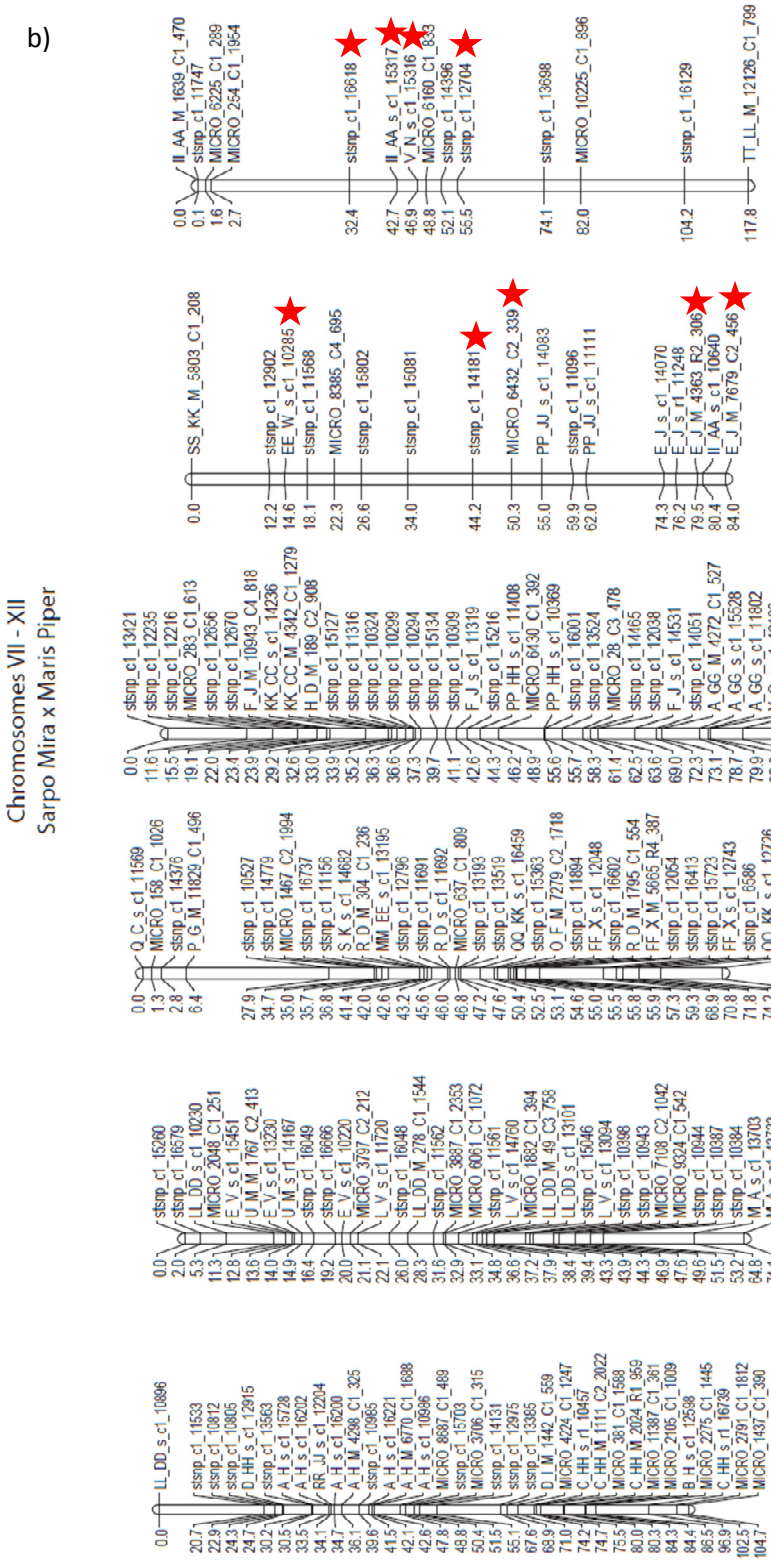


Figure 22 continued.

Chromosome 2

The regression analysis was carried out by Dr Christine Hackett on the phenotypic data and linkage map, and resulted in an output as shown in Table 13.

Table 13 Regression analysis of linked markers on chromosome 2 showing the ID number and name of each marker. Next to the chromosome, the position in cM is shown, as well as the statistical probability levels of the regression analysis for significance and percentage resistance explained by the marker.

No	name	chromosome	cM	significance	explaining
43	stsnp_c1_12294	2	0	0.0348	2.70%
58	Y_P_s_c1_14292	2	28.7	0.01596	3.10%
60	Y_P_s_r1_13931	2	30.7	0.02304	2.70%
62	MICRO_141_C1_802	2	33	0.00834	5.30%
68	stsnp_c1_14826	2	42.9	0.06599	-
82	stsnp_c1_10594	2	75.5	0.0346	2.70%

The regression analysis identified five polymorphic markers with $p < 0.05$ on chromosome 2. The largest combined effect is conferred by the simplex markers [58] (p=0.016) and [60] (p=0.023), which are linked in coupling in Maris Piper. Both (Y_P allele) are located together with marker [MICRO_141_C1_802] (p=0.008) between 28 and 31cM, or spanning a region from 50.5Mb [58] to 56.9Mb [MICRO_141_C1_802] around the two CNL genes DMG 0008306 (51.35Mb) and 0008292 (51.4Mb). Interestingly, a Peroxidase, an enzyme type that was described by Orlowska et al. (2011) to be differentially regulated during the resistance response in Sarpo Mira, resides at position 56Mb. The Y_P allele is linked in Maris Piper in coupling to [MICRO_141_C1_802], and the locus confers a negative resistance response which explains 11.1% of the variation.

Chromosome 3

Three markers from chromosome 3 were identified with an effect on the resistance response. The corresponding physical positions on the DM chromosome are reflected in the order of simplex SNPs [93] (29cM) and [110] (81cM); however the position of SNP [113] at 125cM is different to DM, where it is located at the other end of the

chromosome. Marker [113] is simplex from Maris Piper and explains 5.5% of the resistance response.

To analyse the distribution of genotypes with high and low resistance scores, a variable 'RES' with a value of 1 was introduced for all individual plants with a late blight score of 7 or higher and a value 0 for scores below 7 on Malcolmson's 1-9 scale. The distribution of 0 and 1 was then analysed in a contingency table, to test whether resistance is associated with a respective genotype, and the significance levels were analysed by a Pearson's Chi-square test (Dr Christine Hackett). This test identified a fourth marker SNP [103] at 59.9cM through the contingency table shown in Table 14.

Table 14 Contingency tables were used to analyse the distribution of genotypes with high and low resistance scores. A value RES of 0 (susceptible scores 1-6) and 1 (resistant scores 7-9) was introduced.

RES	0	1
nulliplex	42	1
simplex	55	15
duplex	37	3

The following Chi-square test for association between SNP[103] and RES produced a χ^2 value of 10.14 with 2 degrees of freedom at a probability level p = 0.006. The marker [103] was among those for which no chromosomal position on DM could be retrieved. Although no NB-LRR gene is in the proximity to any of these markers, the NB-LRR targeting miRNA MIR482a (Li et al. 2012) is located between markers [103] and [110], at the physical positions of 35 and 40Mb, respectively. All responses with high significance are from Maris Piper, and the presence of a miRNA, that potentially down-regulates NB-LRRs, within this parent could potentially impact negatively on a resistance response coming from Sarpo Mira.

Chromosome 4

The regression analysis identified six markers around the CNL cluster 29 (chapter 3) and the DMG 0007999 on chromosome 4. The two most significant markers SNP [151] and SNP [154] are duplex markers linked in repulsion in Maris Piper. Together they explain 11.1% of the response, coinciding with a slightly higher resistance score at lower dosages for marker SNP [151]. Marker [154] was also identified in the Chi-square test

(0.01 < p < 0.05) (Table 15). All responses from this locus are linked with lower resistance scores, and thus potentially associated with susceptibility factors that are present in Maris Piper, but not in Sarpo Mira. Comparison of the linkage map and the DM chromosomes showed inverted positions for markers [151], [153] and [154].

Chromosome 5

Only one marker was identified to be highly significant on chromosome 5, SNP [165] (chi-square value 10.92, p < 0.001) (Table 15). This simplex marker from Sarpo Mira was identified through the Chi-square test, and its position around 4.7Mb on the DM genome coincides with the *R1* locus between clusters C33 (4.6Mb) and C34 (4.9Mb) (Figure 23; see chapter 2 and 3). This locus is also known to harbour a large-effect QTL for late plant maturity (Bormann et al. 2004;Gebhardt et al. 2004) leading to physiological resistance at later life stages of the plant.

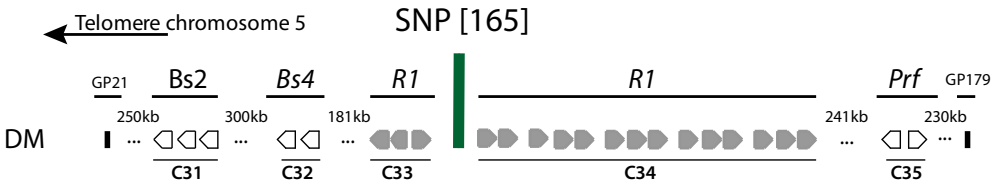


Figure 23 Analysis of the association between polymorphic markers and high resistance levels against the *P. infestans* isolate 3928A identified the SNP marker [165] with a high significance level. This simplex marker is located on chromosome 5, between the *R1* sequence clusters c33 and c34. As shown in the literature, this location coincides with late maturity and elevated late blight resistance.

Chromosome 6

Four markers from chromosome 6 were identified to be related to a significant change in the resistance score (Table 15). All markers are double-simplex linked in coupling for both parents and are located between the CNL-*R* genes 0016656 and 06NN31. Because this locus is not specific to Sarpo Mira and the region is different to DM, it was not followed up.

Chromosome 7

SNP [235] is a simplex marker from Sarpo Mira and is associated with a decrease in the resistance score. This marker is positioned at 47.201 Mb between the TIR-NB-LRR genes 17317 and 07NN5. However, no further marker was identified on chromosome 7 and this marker only explained 1.8% of the variation.

Chromosome 8

The regression analysis identified SNP [248] ($p = 0.035$), and the Chi-square test the simplex marker LL_DD_M_278_C1_1544 on chromosome 8 (Table 15) as significant. These markers are 25cM apart and the locations on DM (2.1Mb and 31.9Mb, respectively) do not coincide with either a NB-LRR position, or another known resistance or susceptibility factor.

Chromosome 9

On chromosome 9 one marker was identified to be linked to a highly significant ($p = 0.004$) change in resistance response, explaining 6.6% of the total response (Table 15). This marker, SNP [282], is duplex by simplex, and resistance levels are slightly higher for a conformation where this marker is either absent or duplex. Because there is no NB-LRR gene identified in the proximity of this marker, an (un)known resistance repression factor might be present, with only one copy in Sarpo Mira.

Chromosome 10

Four significant markers were identified on chromosome 10, of which one (MICRO_1906_C1_431) was 'forced' into its position by JoinMap. However, the DM locations are conserved for the positions of the three mapped markers, supporting the forced position of the fourth marker. The simplex markers SNP [328] and SNP [330] (both PP_HH allele) from Maris Piper, and marker SNP [318] (H_D_M allele) from Sarpo Mira, are associated with an increase in the resistance score. Overall however, effects of both parents are small (MP 4.4%; SM 2.4%). SNP [328] and SNP [330] are located between the NB-LRRs 10NN10 and 0008257, and SNP [318] (47.6Mb) within the large CNL-7 cluster. Additionally, at position 48.9Mb, the NB-LRR targeting miRNA MIR6027 (Li

et al. 2012) was identified. Impact on the resistance is however minor and the role of the miRNA in Sarpo Mira's resistance remains elusive.

Chromosome 11

The regression analysis pointed towards three loci on chromosome 11 that could impact on resistance. The first locus resides on the proximal end and harbours two significant markers (markers SNP [359] and SNP [357], both E_J allele). The other two loci are found at the distal end and are represented by the markers SNP [350], SNP [351], SNP [386] and SNP [346] (the latter two as EE_W allele), see Figure 24. While markers SNP [350] and SNP [351] are the only ones present in Maris Piper (duplex), all other alleles are simplex in Sarpo Mira and linked in coupling (Table 15).

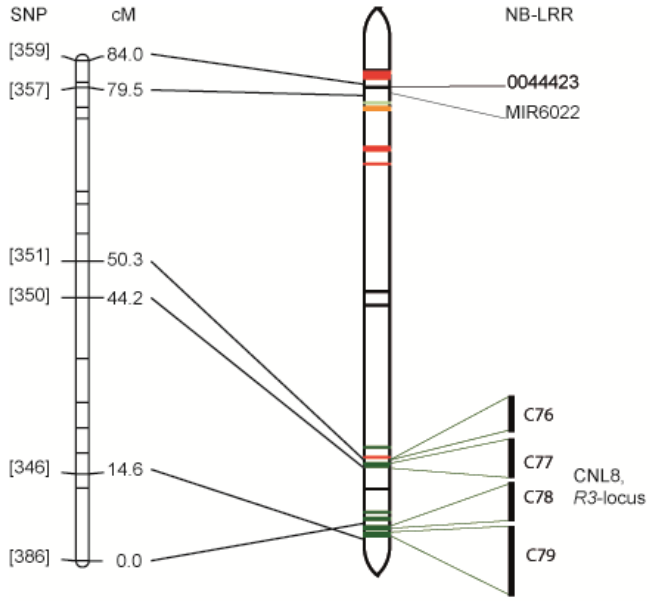


Figure 24 Comparison of the SM×MP linkage map based on polymorphic markers and the chromosomal map of DM NB-LRRs from chapter 2 and 3. Highlighted are three loci on chromosome 11 with significant effect on the resistance response. In Sarpo Mira, all three loci are simplex and linked in coupling, while SNPs [350] and [351] are also present in Maris Piper (duplex). Lines between the maps connect the regions and show synteny for the two top loci, and an inversion for the lower loci. Bars to the right indicate large NB-LRR cluster, all belonging to the CNL-8 group of *R3*-like genes.

The E_J allele is simplex from Sarpo Mira and is linked to an increase in resistance, accounting for 5.1% of the variation in total. Furthermore, marker SNP [359] was identified in the Chi-square test (Pearson Chi-square value is 8.20) to be significantly associated with a higher resistance score ($p = 0.004$). This allele with positions at 3.8Mb

and 4.7Mb is flanking the CNL 0044423 at 4.2Mb, as well as the NB-LRR targeting miRNA MIR6022 at 4.3Mb (Figure 24).

The markers SNP [350] and SNP [351] explain 12% of the resistance. The duplex by simplex SNP [351] is linked in Sarpo Mira to the alleles E_J and EE_W. Both markers, at positions 34.3Mb and 35.9Mb, flank the CNL-8 *R3-like* NB-LRR 0007344 (35.4Mb) (Figure 24). The Sarpo Mira specific EE_W allele is simplex and both markers together explain 5.9% of the resistance. As shown in Figure 24, a large stretch of the *R3* locus between 40.2 and 41.8Mb is flanked by these markers. Sarpo Mira and genotypes with all three alleles present have the highest resistance, and all markers of this chromosome can explain 22.4% of the total resistance response (summarised in Table 15).

Chromosome 12

A significant effect on the resistance was found for four very closely linked markers on chromosome 12. Of those two SNPs [365] and [366], are simplex markers from Sarpo Mira and linked in repulsion. Marker SNP [365] (II_AA allele) is associated with a higher value for resistance, and was also identified in the Chi-square test ($0.01 < p < 0.05$). SNP [366] (V_N allele) is however associated with a lower value for resistance, and SNP [364] is a double-simplex marker linked in coupling to II_AA in Sarpo Mira. SNP [369] is a duplex by simplex marker linked to V_N in coupling in Sarpo Mira, and if present in simplex has a higher resistance score. All results are summarised in Table 15.

The Sarpo Mira derived markers can explain 15% of the resistance. The linkage position is distorted, when compared to the physical positions on the DM chromosome as can be seen on the position of the last marker [Micro_12126_C1_799] in Figure 25. However, when the DM derived positions are used, all four markers are situated within 3.6Mb and flank a single CNL-R 1007575, which has sequence similarity to the *Arabidopsis* NB-LRR *RPS2*.

In summary, this analysis has identified several genetic loci that explain mainly relatively small percentages of the resistance/susceptibility in Sarpo Mira or Maris Piper. This is however in accordance with the high complexity of the resistance seen also in the segregation pattern of the SM×MP population. It was interesting to see that many effects coming from Maris Piper have a negative influence on the resistance. Some larger effects were identified on chromosome 5 within the *R1*/maturity locus, on

chromosome 11 where three loci are linked in Sarpo Mira that flank the *R3* locus and on chromosome 12 around a CNL-R gene. The loci on chromosome 11 and 12 together explain 37.4% of the total resistance response towards the isolate 3928A.

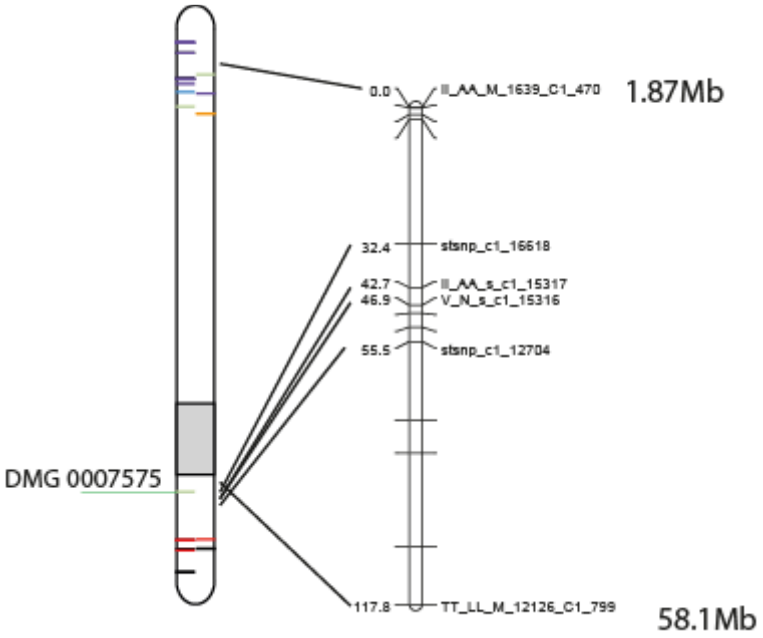


Figure 26 Comparison of the SM×MP linkage map based on polymorphic markers and the chromosomal map of DM NB-LRRs from chapter 2 and 3. Highlighted is one locus on chromosome 12 with significant effect on the resistance response. Lines between the maps connect the regions and show synteny and distortion of the maps, anchored using the first and last SM×MP SNP marker at cM position 0.0 and 117.8.

Table 15 Summary of all results gained by the regression analysis and Chi-squared test.

					Regression analysis		Chi-squared test		Dosage in parent		
No.	name	LG	cM	DM pseudomolecule v3 2.1.10	F Test probability	% explaining response	p-value	X^2	Maris Piper	Sarpo Mira	effect from
58	Y_P_s_c1_14292	2	28.7	50509289	0.01596	3.1			simplex	-	
60	Y_P_s_r1_13931	2	30.7	51313791	0.02304	2.7			simplex	-	MP
62	MICRO_141_C1_802	2	33	55318694	0.008	5.3			simplex	duplex	
93	G_D_s_c1_15342	3	29.6		0.032	2.4			-	simplex	
103	stsnp_c1_16513	3	59.9		-	-	0.006	10.1	simplex	simplex	MP
110	stsnp_c1_11927	3	81.3		0.041	3.0			simplex	duplex	
113	JJ_BB_M_12033_C1_1104	3	125.8		0.001	5.5			simplex	-	
145	stsnp_c1_11669	4	57.1	55593537	0.057	-	<0.05	-	triplex	simplex	
151	stsnp_c1_10422	4	68.9	60117268	0.00184	5.9			duplex		MP
153	stsnp_c1_15240	4	70.8	59748507	0.36619	-			-	-	
154	stsnp_c1_11463	4	71.8	58634758	0.00141	6.2			duplex	-	
165	JJ_BB_Ctg151_136	5	10.4		-	-	<0.001	10.9	-	simplex	SM
194	stsnp_c1_13878	6	22.9		0.00393	5.9			simplex	simplex	
195	stsnp_c1_10901	6	25.3		0.0286	2.9			simplex	simplex	both
198	stsnp_c1_10559	6	26.5		0.03588	2.7			simplex	simplex	
	MICRO_669_C3_729	6	*		0.00676	4.9			simplex	simplex	
235	C_HH_s_r1_10457	7	53.6		0.04311	1.8			-	simplex	SM

Table 15 Continued.

248	stsnp_c1_15260	8	5.6		0.03571	3.1			simplex	simplex	both
-	LL_DD_M_278_C1_1544	8	30		-	-	<0.05	-	simplex	-	
282	stsnp_c1_16737	9	35.7	13792203	0.00409	6.6			duplex	simplex	both
318	H_D_M_189_C2_908	10	33	47556298	0.02573	2.4			-	simplex	
328	PP_HH_s_c1_11408	10	46.2	26991674	0.03116	2.1			simplex	-	both
330	PP_HH_s_c1_10369	10	55.6	26358916	0.02558	2.3			simplex	-	
-	MICRO_1906_C1_431	10	*	1296439	0.01752	3.7			simplex	simplex	
386	EE_W_s_r1_15798	11	*	40157740	0.016	3.1			-	simplex	
346	EE_W_s_c1_10285	11	15.5	41832218	0.01531	2.8			-	simplex	
350	stsnp_c1_14181	11	45.3	34264732	0.01083	5.3			duplex	simplex	SM
351	MICRO_6432_C2_339	11	51.4	35977124	0.00269	6.7			duplex	simplex	
357	E_J_M_4363_R2_306	11	80.5	4723853	0.02933	2.3			-	simplex	
359	E_J_M_7679_C2_456	11	85	3768964	0.01823	2.8	0.004	8.2	-	simplex	
364	stsnp_c1_16618	12	32.4	55877358	0.04705	2.7			simplex	simplex	
365	II_AA_s_c1_15317	12	42.7	58687209	0.00177	5.7	<0.05	-	-	simplex	SM
366	V_N_s_c1_15316	12	46.9	59441338	0.03128	2.4			-	simplex	
369	stsnp_c1_12704	12	55.5	56286460	0.01564	4.2			duplex	simplex	

4.2.6 Enrichment of NB-LRR specific sequences from resistant and susceptible bulks of Sarpo Mira and Maris Piper

In chapter 3 the successful enrichment of NB-LRR specific sequences from potato gDNA prior to Illumina sequencing is reported. Here, this sequence capture was applied to sequence the NB-LRR gene complement from the previously defined bulks of SM×MP accessions that are resistant and susceptible to 3928A. The chief driver for this analysis was to identify candidate NB-LRRs or informative polymorphisms within NB-LRRs that could explain the resistance.

Similar to the results in chapter 3, the enrichment efficiency was determined by qPCR using NB-LRR specific primers, and is detailed in Table 16. The log phase of amplification with *R3a* specific primers was reached 9.0 cycles earlier for NB-LRR enriched bulk resistant (BR) samples, and 8.3 cycles for the bulk susceptible (BS) samples if compared to non-enriched samples. Similarly, *R2* specific primers amplified 9.2 and 8.3 cycles earlier from enriched BR and BS samples, respectively, if compared to non-enriched controls. Amplification of the endogenous ribosomal *18S* gene showed only minor differences between enriched and non-enriched samples. A t test identified statistically significant differences between the mean cycles to threshold values for enriched and non-enriched samples, with exception for *18S* BR, with a P value of 0.0980.

Illumina paired-end sequencing yielded 82.3 million raw reads for BR and 70.1 million for BS (Table 18). Quality control removed 16% and 41% reads, respectively, as these contained ambiguous nucleotides (N's) in their sequence.

Mapping of paired-end reads using Bowtie (allowing for four nucleotides mismatch in the extension) assembled 60 million BR- and 36 million BS-reads to the DM superscaffolds (DMB), which is in both cases 88% of the quality controlled reads (Table 17). Furthermore, 41% of the BR reads and 38% of the BS reads could be mapped to the 762 DM NB-LRRs.

The aim of the Illumina sequence analysis was to compare SNP ratios over NB-LRR genes between BR and BS, for which we expect 25% of the resistance allele in BR and less than 10% in BS, due to the tetraploid background. All quality controlled paired-end Illumina reads were therefor *de novo* assembled using Velvet (embedded in Galaxy). In total 59,366 contigs were produced with an N50 of 191 and a maximum contig size of 17kb. Attempts to scaffold these contigs based on the DM reference failed, due to high sequence conservation between the alleles of a single gene. It was decided to leave this data until better assembly packages became available.

Table 16 qPCR was carried out to determine the efficiency for enrichment of NB-LRR specific fragments from bulked resistant (BR) and susceptible (BS) gDNA. C(t) values are shown for each replicate, and the mean that was used to calculate the ΔC(t) as difference between enriched and non-enriched templates. ΔΔC(t) values represent the actual difference after subtracting the ΔC(t) from the control *18S*. P-value was calculated for the difference between enriched and non-enriched samples using a *t* test.

Target	template		mean	Stdev	ΔC(t)	ΔΔC(t)	P
R3a	BR	enriched	20.19	0.06	9.9	9.0	< 0.0001
		non-enriched	30.07	0.19			
	BS	enriched	19.23	0.06	9.7	8.3	< 0.0001
		non-enriched	28.99	0.05			
R2	BR	enriched	19.73	0.21	10.1	9.2	0.0002
		non-enriched	29.82	1.37			
	BS	enriched	18.31	0.07	9.7	8.3	< 0.0001
		non-enriched	28.09	0.19			
18S	BR	enriched	21.09	0.76	0.9	-	0.0980
		non-enriched	22.04	0.09			
	BS	enriched	20.48	0.27	1.4	-	0.0008
		non-enriched	21.91	0.04			

Table 17 Quality control and initial analysis of the paired-end Illumina reads for both bulks of SM×MP. The analyses were carried out using scripts embedded in the TSL Galaxy instance. Illumina quality scores were determined for each position and are shown in the table as lowest mean quality score. Reads were paired-end mapped using Bowtie to the DM superscaffolds (DMB) and to the 762 DM NB-LRR sequences.

	BR	%	BS	%
file size (each end)	8.8Gb		7.5Gb	
total reads	82,376,758		70,130,526	
seqs with N removed	13,370,562	16	28,637,406	41
reads remain	69,006,196		41,493,120	
lowest mean quality score	36		36	
mapping to DMB	60,800,383	88	36,373,775	88
mapping to DM NB-LRR	28,009,662	41	15,733,637	38

Furthermore, the data was also analysed by Kare Lehman Nielsen (Aalborg University, Section of Biotechnology). In his lab the data was mapped to the reference potato genome sequence, and based on that SNPs were detected which served as markers for QTL regions harbouring significantly different NB-LRR genes. The coverage distribution was assumed to be non-random at these QTL positions, and tested with Fisher's exact test. As an outcome, two candidate NB-LRR genes were identified on chromosome 7 (DMG 1030700) and 8 (DMG 0002277) with a SNP ratio of around 90% in BS and 25% in BR. However, these positions did not correspond to the loci, previously identified by the genotyping analysis.

4.3 Discussion

The potato cultivar Sarpo Mira is known for its good resistance against contemporary late blight isolates. Unravelling the resistance factors that are present in this cultivar would enable breeders to use it more efficiently in breeding programmes for new blight resistant cultivars. In this multi-disciplinary study, it was anticipated to shed more light on the resistance response of Sarpo Mira against three aggressive late blight isolates of the MLG 6_A1 and 13_A2 that occur in high abundance in the UK and Western Europe. Initial resistance screens with Sarpo Mira involved detached leaf assays showed overall susceptibility, as previously described by Rietman et al. (2012) when Sarpo Mira was tested with the *P. infestans* isolate IPO-C. This was however in stark contrast to the strong resistance of Sarpo Mira reported from various field trials against these isolates (White and Shaw 2009). Therefore late blight resistance screens were carried out as whole plant assays as described in Colon et al. (2004). The response seen in this study towards 13_A2 isolate 3928A was not absolute resistance, but rather a slow disease development that did neither infect the whole plant nor ended in a finished life cycle of the pathogen. This and the observed segregation pattern of 1:7:1 (resistant:moderately susceptible:susceptible) suggests that the resistance is a "field resistance" conferred by potentially three or even more genetic factors, similar to that seen earlier in the potato cultivar Stirling (Bradshaw et al. 2004). Resistance to 7454A was however absolute and segregated close to 1:1, suggesting one major *R* gene to be involved in this particular response.

Both whole plant assays indicated that (most) plants resistant to 3928A are also resistant to 7454A, but not *vice versa*. Only two clones showed a specific resistance to 3928A (though with lower resistance scores), but six were specifically resistant towards 7454A. The major *R* gene in action against 7454A may thus potentially play also a role in the resistance towards 3928A. As the resistance response towards 3928A is a delayed susceptibility, it can be hypothesized that these "additional" genes are rather resistance factors than major *R* genes. They could for example encode for miRNAs that control NB-LRR transcription or other plant components (Li et al. 2012;Orlowska et al. 2012). This is in agreement with the statement of Rietman et al. (2012) that "the genetic basis of late blight resistance in Sarpo Mira is highly complex".

The availability of several *P. infestans* genome sequences to date (Haas et al. 2010;Cooke et al. 2012) allowed a prediction of their RXLR effector gene complement, which has been exploited in several studies to identify recognition specificities in Solanaceaes (Rietman et al. 2012;Vleeshouwers et al. 2011). Rietman et al. (2012)

screened Sarpo Mira plants for the recognition of 234 predicted RXLR effectors from the *P. infestans* isolate T30-4 in an effectoromics-termed approach. This study showed that Sarpo Mira recognizes five different effectors, and is thus an example for a cultivated variety with functionally stacked resistance genes that has not yet been broken by contemporary late blight isolates. Among the five recognized effectors are *Avr3a, Avr3b, Avr4*, as well as two novel avirulence genes termed *AvrSmira1* and *AvrSmira2* (Rietman et al. 2012).

Comparison of results obtained by Rietman et al. (2012) with gene expression analysis of 3928A conducted by Cooke et al. (2012) and our own data identified *AvrSmira1* as potential effector candidates. *AvrSmira2* is another effector candidate, as expression analysis in Cooke et al. (2012) focused on 2 and 3 dpi, while Rietman et al. (2012) showed its expression at 17 hpi. Rietman et al. (2012) do not mention the origin of this expression data, or how the analysis was carried out. The remaining two effectors were determined to be either absent, in case of *Avr3b*, or truncated, in case of *Avr4* (van Poppel et al. 2009).

The sequences of the RXLR effectors that were screened by Rietman et al. (2012) on Sarpo Mira, although synthesized, where solely derived from the isolate T30-4. Cooke et al. (2012) have shown that 3928A has six RXLR effector candidates that are absent from T30-4, but which are also present in MLG 6_A1. Screening these six effectors might reveal another avirulence gene, recognized by Sarpo Mira, and thus an *R* gene that is responsible for resistance towards both 3928A and 7454A.

The delayed onset of blight symptoms in Sarpo Mira is similar to reports from plants containing *R8* (Kim et al. 2012). The tested blight isolate 3928A was also shown to be avirulent on *R8* containing plants (Cooke et al. 2012). However, *R8* is unlikely the sole resistance determinant as isolate 7454A is virulent on *R8*, and in addition no polymorphic marker was identified in the proximity to the recently re-mapped *R8* locus on chromosome 9. A function of *R8* as one of the previously discussed additional resistance factors can however not entirely be excluded. In fact only recently, *Avr8* was determined to be *AvrSmira2* (Dr Jack Vossen personal communication).

The Genotyping assay had sufficient markers to cover the linkage groups by length however the density of markers was not sufficient across all of the linkage groups. The small size of this tetraploid mapping population allowed only the detection of QTLs explaining a small percentage of the resistance variance (Dr Christine Hackett personal communication). For a more thorough QTL analysis the population size would need to

be increased, as well as the number of markers used, to define a much denser map. Interestingly the biggest effects were determined from the least marker dense chromosomes 11 and 12. It can be hypothesised that these chromosomes from Sarpo Mira are too diverse in comparison to the cultivars that were used to create the markers.

Polymorphic markers with positive effect on resistance coming from Sarpo Mira were determined within the *R1/*maturity locus on chromosome 5, on chromosome 11 close to the *R3* locus and on chromosome 12 around a CNL-R gene of unknown function. Sarpo Mira is known to be a late maturing variety (www.europotato.org), and therefore it was not surprising to find a marker very close to the QTL for late maturity within the *R1* locus. This provides also a hint towards *S. demissum* as a potential progenitor for Sarpo Mira (Gebhardt et al. 2004). Although *R1* has been overcome by contemporary field isolates of *P. infestans*, the surrounding genome segment originating from *S. demissum* still contributes to quantitative resistance (Gebhardt et al. 2004). This thesis reports on the identification of 17 *R1*-family members in this region, and thus the assumptions by Gebhardt et al. (2004) that more NB-LRR family genes are located in this region are true. This might point towards a class of NB-LRR genes with minor effects on resistance.

The capture of NB-LRR specific sequences yielded similar results as in chapter 3 described for DM. This suggests that the hybridisation was successful although the genome of the *S. tuberosum* Group Tuberosum variety Sarpo Mira is not very distinct from the *S. tuberosum* Group Phureja clone DM or that indeed the probes allow for significant diversity as suggested in chapter 3. However, early attempts of using de novo assembly to re-establish whole genes or large parts of them failed. Because of the tetraploidy of Sarpo Mira, the number of NB-LRR genes will be four-times that of the reference DM. Therefore subfamilies like R3 with over 50 highly similar sequences in DM (Chapter 4) are likely to have around 200 members in Sarpo Mira. Assembly of these short 76bp reads and scaffolding of contigs to these highly similar sequences is thus very difficult (Chapter 5). The approach to identify SNPs between BR and BS reads, used in the lab of Kare Lehman Nielsen identified two loci that were not identified during the genotyping in this study. Here, the incomplete NB-LRR set derived from the PGSC annotations was used to map the Illumina reads to and identify SNPs in a ratio of less than 5% in BS and around 25% in BR. It was further tested whether a very close sequence homolog to one of both NB-LRRs is in the proximity to any of the identified loci linked to the resistance. However this gave no positive result. The genetic and

genomic background of the resistance towards 3928A therefore remains therefore elusive.

4.4 Material and Methods

4.4.1 Crosses and Plant material for whole plant assay and genotyping

The resistant *S. tuberosum* Group Tuberosum cultivar Sarpo Mira was crossed in 2006 as female parent with the susceptible cultivar Maris Piper (John Bradshaw, Geoff Swan). Seeds were sown in 2007 in a glasshouse, 230 seedlings were raised and tubers harvested to be maintained yearly in the glasshouse as well as in the field at the James Hutton Institute (clonal maintenance done by Geoff Swan). Glasshouse grown plants were used for disease testing in 2010, 2011 and 2012. The plants were grown single-stemmed from tubers and young leaflets were sampled, frozen in liquid nitrogen and stored at -80C for the genotyping experiment. Genomic DNA from flash frozen leaf tissue was extracted using the Qiagen DNeasy kit (Qiagen), according to manufacturers protocol and stored at -20C.

4.4.2 *Phytophthora infestans* isolates and disease testing

Phytophthora infestans MLG 13_A2 isolate 06_3928A, MLG 6_A1 isolate 7822B and 7454A were kindly provided by Dr David Cooke (JHI) and weekly maintained on detached leaves of *S. tuberosum* cv. Craig's Royal for at least three weeks prior to inoculation assays.

A detached leaf assay was carried out on leaflets of Sarpo Mira and Maris Piper as described in Vleeshouwers et al. (1999) with minor modifications. Leaflets were directly placed onto wet tissue with the abaxial side up and one 20µl droplet of a zoospore suspension with 17,000 sporangia/ml, was applied onto both sides of each leaf. Boxes were closed to maintain high humidity and stored under light conditions at 16°C. Developing symptoms were observed daily for eight days.

Two plants for each of the 154 clones, and four of each parent were tested in a whole-plant glasshouse test, modified from Colon et al. (2004). Plants were ready for testing when the first flower buds appeared within the canopy. A zoospore suspension was adjusted to 17,000 sporangia/ml, incubated for up to 3h at 4°C, and sprayed over the plants to the point of runoff, as described by Colon et al 2004, using a commercial hand-

pump sprayer. After incubation for 18h darkness at 90 - 100% relative humidity, plants were moved to a north facing glasshouse with a temperature of around 16°C. Each replicate was scored 8 dpi on Malcolmson's 1–9 scale of increasing resistance which is related to the percentage of necrotic tissue as illustrated by Cruickshank et al. (1982) (Annex 3). Results were assumed to be consistent if the score was within one scale step for both plants.

Table 18 PCR primers used to amplify the full or partial (qPCR) effector gene from gDNA.

FJ#	name	target and remarks	5' - 3' sequence
194	Avr3a full F	PITG_14371	ATGCGTCTGGCAATTATGCTGTC
195	Avr3a full R		CTAATATCCAGTGAGCCCCAGGT
196	Avr3a qPCR F		AGAGCAGATGCCAAAAAGCTAGC
197	Avr3a qPCR R		GGTCTAGCGTAACCCTGTTGTGC
183	Avr3b full F	PITG_18215	ATGCGAGCCTACTTTGTCCT
184	Avr3b full R		TTAGAAATTGTTCTTTCCGGTC
185	Avr3b qPCR F		CGGCTTAACCAAGGAATGGACGT
186	Avr3b qPCR R		TCACGAGAGCGTCCCAGTTCTG
190	Avr4 full F	PITG_07387	ATGCGTTCGCTTCACATTTTGCT
191	Avr4 full R		CTAAGATATGGGCCGTCTAGCTTG
192	Avr4 qPCR F		AACCCGCAGCAGTATGCCAAGTTC
193	Avr4 qPCR F		CTTTTTATATGGGTTCCCTCCATGG
187	AvrSmira1 full F	PITG_07550	ATGCGTCTAAGCTCCACATTTCTTG
188	AvrSmira1 full R		TTATCCGGAGGGGTTTAGCGAGT
181	AvrSmira1 qPCR F		ACCCCGGTCAACAAGAAGGCCT
182	AvrSmira1 qPCR R		
189	AvrSmira2 full F	PITG_07558	ATGCGCTCAATCCAACTTCTG
155	AvrSmira2 qPCR F		GGCTAGTGCACGTAGGGACG
127	AvrSmira2_ATG_F	no signal P, ATG insert	AAAAAGCAGGCTTCATGACACCAGCACCGCCACAAGT
146	AvrSmira2_attb_R		GGGGACCACTTTGTACAAGAAAGCTGGGTCTTACGAT GTTTTCGCTTCTTTAAAAAGC

4.4.3 Determination of candidate effector presence

A highly concentrated suspension of sporangia and zoospores was made for the three *P. infestans* isolates 3928A, 7454A and 7822B by rinsing infected leaf material 7dpi with ice cold distilled water. The suspension was subsequently centrifuged at 3,000 rpm at 7°C for 5 min, the supernatant was discarded and the pellet flash frozen in liquid nitrogen and stored at -80°C. Genomic DNA was extracted using the Qiagen DNeasy kit (Qiagen),

according to manufacturer's protocol. The gDNA was then stored at -20°C. GoTaq polymerase amplifications were carried out on gDNA using the primers disclosed in Table 18. As a control a primer pair for the house-keeping gene Actin was used. PCR reactions were carried out in a volume of 10µl (1x GoTaq Buffer, 0.5mM dNTPs, 0.5mM primers) in a thermal cycler at 95°C for 5 min, cycling at 95°C for 25s, 60°C for 25 s, and 72°C for 15s, and a final 72°C for 5 min.

4.4.4 Genotyping and linkage analysis

Genotyping of 176 individual F1 plants plus the parents of the SM×MP population in duplicate was accomplished with the GoldenGate assay (Illumina). The experiment was carried out by the Sequencing and Microarray facility at the JHI. SNP markers used in this experiment are organized in potato oligo pooled assay (POPA)-libraries 1, 2, and 5, all identified from alignments of EST sequences of three US potato varieties with the DM genome sequence (Dr Sanjeev K. Sharma). Chromosomal positions for each SNP marker were established using local megablast searches against PGSC_DM_v3_2.1.10_pseudomolecules.fasta.

The linkage map was constructed from the SNP theta scores derived from the GoldenGate assay, using a method and computer programmes developed by Dr Christine Hackett (Biomathematics and Statistics Scotland), who carried out most of this analysis. A paper describing the method in full has been submitted for publication. In brief, the SNP dosage (as AAAA, AAAB, AABB, ABBB, BBBB) were inferred from the theta scores. After identification of simplex SNPs, their linkage was analysed when linked in coupling phase, carried out using JoinMap (JoinMap 4, Kyazma B.V.). Chi-squared tests of independence of these simplex SNPs linked to duplex and to double-simplex SNPs aided the tentative identification of homologous chromosomes. These tentative groups were then used as a framework to cluster SNPs of all dosages into linkage groups. Within each of the derived 12 linkage groups, recombination frequencies and lod scores are calculated between all pairs of SNPs and the most likely phase is determined. Each set of pairwise data is ordered using JoinMap, and the overall phases are established.

A regression analysis of the resistance scores to the mapped markers was carried out, reporting all effects with $p < 0.05$, ignoring multiple testing effects. In a further test, the association was examined between resistance and genotype. Therefore, a variable RES was introduced with a value 1, for a resistance score of 7 or higher, or a value 0 for

scores below 7. The data was displayed in a contingency table, and the distribution and significance levels were analysed using a Pearson's chi-squared test.

The positions of significant markers were compared to the positions of the previously established DM NB-LRR positions and the literature.

4.4.5 Enrichment, Illumina sequencing and Sequence analysis

Previously extracted DNA of the 17 most resistant (SM×MP: 14, 53, 89, 90, 93, 94, 100, 104, 118, 123, 141, 160, 165, 166, 178, 180, 221) and 19 susceptible (SM×MP: 1, 3, 4, 7, 15, 21, 23, 42, 61, 78, 88, 98, 120, 187, 194, 196, 197, 198, 202) individuals was combined in equimolar ratios of 2µg each into BR and BS samples. A total of 40µg DNA per sample was sheared in a Covaris sonicator for 30s (20% duty cycle, intensity level 5, 200 cycles/burst). Sheared DNA was further processed as described in chapter 3.4.2.

Enrichment efficacy was assessed in a quantitative PCR SYBR Green assay (Sigma-Aldrich Co.) with 1ng starting material from enriched and non-enriched samples, 0.5µl of the corresponding NB-ARC specific primer (10mM) and 8µl water. Primer information is disclosed in Table 10. Mean C(t) values were compared between enriched and non-enriched samples using a *t* test software, available at http://www.graphpad.com/quickcalcs/ttest2/.

The enriched DNA libraries were paired-end sequenced (read length 76bp) on the Illumina GAII at the Sainsbury Laboratory. Raw sequence data processing was carried out as described in chapter 3.4.3.

Chapter 5 - General Discussion

The NB-LRR gene family in the potato genome annotation

The potato genome provides an unprecedented opportunity to establish the physical position, distribution and phylogenetic relationship between NB-LRR-type resistance genes. These genes mediate arguably the most important form of inducible resistance towards adapted pathogens and lead to effector triggered immunity by directly or indirectly perceiving pathogen effector molecules (Jones and Dangl 2006). The genome blueprint formed the basis of this thesis which was aimed to devise a strategy to exploit the potato genome to more rapidly map and ultimately clone novel resistances. In chapter 2, we developed a pipeline to aid the identification of the NB-LRR genes from the published potato gene models. In total, 438 NB-LRR genes were identified and the study was peer reviewed and published in BMC Genomics (Jupe et al. 2012). In chapter 3, this knowledge was transferred into a NB-LRR enrichment tool comprising over 49.000 biotinylated RNA baits. In a proof of concept study, the efficacy of these probes to capture and enrich NB-LRR genes prior to Illumina sequencing of genomic DNA fragments was established by using the sequenced potato clone DM. Out of the 438 initially predicted NB-LRR genes 424 were verified, but further 338 NB-LRR encoding sequences were discovered. These results showed that 44% of the final NB-LRR gene complement was derived from previously un-annotated regions of the potato genome. This highlights problems of automatic gene calling algorithms during genome annotations. Most new NB-LRRs are located within or around previously defined clusters, and possess a high degree of sequence homology as initial analysis has shown. Whether exon structure was different to the previously identified NB-LRRs was not determined. It remains elusive why these genes have not been identified during the annotation process by the PGSC. Further it needs to be investigated whether other gene families are also affected and not fully annotated. The Tomato Genome Consortium (TGC 2012) have made the effort to re-annotate the potato genome using their gene prediction pipeline and identified even fewer protein-coding genes, 35,004 compared to 39,031 by the PGSC (2012). Compared to the PGSC analysis of potato which identified 406 NB-LRRs, the tomato genome consortium derived gene models contained 564 NB or TIR containing genes. Within the tomato genome, so far 309 NB-LRR genes were identified (Andolfo et al. 2013), but the high synteny between the tomato and potato genomes could hint at a higher number, similar to that found in this study in DM. However, to elucidate if the NB-LRR gene predictions in tomato are more accurate compared to the initial NB-LRR gene analysis in potato and therefore to establish if the

tomato genome indeed encodes for fewer NB-LRRs, an enrichment and sequencing approach as described for DM in this thesis would be required. It has become apparent from other genome studies that NB-LRR genes are very difficult to predict. For example, misannotations of the NB-LRR gene family have also been reported in *Arabidopsis*. In 2003, Meyer et al. manually re-annotated the *Arabidopsis* NB-LRR genes and identified 56 wrongly annotated genes. In this study (Jupe et al. 2012) a manual re-annotation of the potato genome increased the number of full length TNL genes by 20, and that of CNLs by 84, compared to the annotations by the PGSC (2012). To fully understand the NB-LRR gene family, their evolution and functions, it is crucial to identify the full complement from sequenced plant genomes. With the rapid technological advances seen in recent years, it is tempting to speculate that a NB-LRR gene specific enrichment and sequencing has the potential to identify additional resistance gene candidates in already sequenced plant species such as *Brassica rapa* and *B. napus*, *Vitis vinifera*, *Populus trichocarpa* and *Oryza sativa*. The NB-LRR enrichment pipeline designed and described in this Thesis has been shown to have the potential to re-annotate, with some modifications of the bait-library composition, the NB-LRR gene complements from all draft plant genomes.

Genome sequences provide evidence for a co-evolution of pathogen and host

The underlying analysis has provided us with an unprecedented insight into the vast expansion of NB-LRR gene clusters throughout the genome. The identified 85 clusters which contain up to 27 closely related members - in case of the chromosome 7 located cluster C17 - provides some evidence for high selection pressure within this important resistance gene family (Hulbert et al. 2001;Michelmore and Meyers 1998;Parniske et al. 1997). Homogenous NB-LRR gene clusters have evolved through duplication events, but the origin of heterogeneous clusters is unclear (Andolfo et al. 2013). Furthermore, CNLs are most likely to be clustered (Andolfo et al. 2013;Jupe et al. 2012).

Regardless the large number of 762 NB-LRR genes, the sequenced potato clone DM is described as a rather fragile plant without known resistances (PGSC, 2012). It can only be speculated about potential functions of this large gene family in, for example non-host resistances (Schulze-Lefert and Panstruga 2011), or as regulators-of-regulators, with respect to miRNA-based regulation of NB-LRR expression (Li et al. 2012;Lozano et al. 2012).

Similarly, genome comparisons of the aggressive *P. infestans* isolate 3928A with the less aggressive T30-4 isolate revealed 320 duplication events of RXLR effector genes, and the deletion of 47, accompanied by gains and losses of gene induction (Cooke et al. 2012;Haas et al. 2009;Raffaele and Kamoun 2012). These results support the hypothesis of a very dynamic *phase 4* of the "zig-zag-zig" model, the co-evolution of host and pathogen (Hein et al. 2009b;McDowell and Simon 2006). Within the potato genome, NB-LRRs have, with a few exceptions, a 'normal' distribution that is not different from the average DM gene (this Thesis; Jupe et al. 2012). *P. infestans* effectors however, are known to reside in gene-sparse but highly repetitive areas of the genome (Haas et al. 2009) in which genome comparisons have identified the most changes (Raffaele and Kamoun 2012). From a different point of view however, the mainly homogeneous clusters in which 85% of the DM NB-LRR genes reside in, represent regions of highly repetitive nature, and are thus aiding higher recombination rates (Anderson et al. 2006). This was shown for the model plant *Medicago*, where the highest NB-LRR gene density coincides with the highest recombination rates (Paape et al. 2012).

Efficiency of gene capture assays is dependent on probe design and sequence analysis

Capture of DNA targets prior to sequencing allows the reduction of genome complexity and helps increasing the throughput and output of informative sequences (Andolfo et al. 2013;Gnirke et al. 2009;Saintenac et al. 2011;Stower 2011). As could be shown in this thesis, only 0.24% of the potato genome encodes for NB-LRR genes. To be able to call nucleotide polymorphisms in comparative analyses from these genes, it is absolutely necessary to reduce the genome complexity. The capture of NB-LRR specific fragments from the reference potato clone DM and bulks of a tetraploid potato population showed similar efficiencies in that 40 to 50% of the Illumina sequence reads are derived from NB-LRR-like genes. As previously discussed in chapter 3, this number is however lower than the 60% of reads on target in a maize study (Saintenac et al. 2011), and the approximately 70% observed in a comparative study of two different *in solution* based human exome capture kits (Sulonen et al. 2011). The success of the enrichment is very much dependent on the design of the bait library, and the efficiency of the method used to analyse the sequence reads (Biesecker et al. 2011;Stitziel et al. 2011). Comparisons between enrichment experiments show a wide range of different bioinformatical approaches used to analyse the data, and lacking a common strategy (Stitziel et al. 2011). Of critical importance to our studies were not only the settings within the read

mapping software but the reference sequence itself that was used to map the reads and thus to determine the enrichment efficiency. The initially anticipated strategy was based on the identification of polymorphisms from re-assembled NB-LRR genes. Sequence analysis provided in chapter 3 and 4 independently showed the difficulty of *de novo* assembly of NB-LRR genes. The main problem is the large number of paralogues and alleles with sequence similarities of 100% over parts of the gene. This was the main reason why this approach failed when tested on the enriched and sequenced bulks of the tetraploid SM×MP population. In DM it only worked because of the exact reference to which reads were mapped. The short Illumina read length obtained (76bp) is under these conditions not suitable to de-convolute alleles from paralogues and to establish the true NB-LRR gene complement, and longer reads are required for future studies. The presented analysis of DM NB-LRR enrichment reads was only possible because DM itself could be used as a mapping reference. All *Rpi* genes so far cloned are derived from diploid species, or those that were made diploid through crosses with other clones, for example *R1* and the *R3* genes (Ballvora et al. 2002;Huang et al. 2004). With increasing number of chromosomes, the number of homologous sequences is rising. Certain clusters, like the *Rpi*-gene cluster *R2* on chromosome 4 and *R3* on chromosome 11, are expected to harbour more than 200 homologs in a tetraploid potato cultivar and the chances of mis-assembly are very high.

The enrichment-based identification of functional NB-LRR genes is dependent on the NB-LRR gene used for the bait-design. In an *in silico* experiment, we were able to show that the baits used in our bait-library are able to pull out gene fragments with 75% or less homology (50% in some cases). Functional genes, such as *Rpi-blb1*, for which the closest homolog found in DM has less than 80% sequence identity, are expected to be pulled out using our bait-library, showing the great potential of this tool. Next to the canonical CNL and TNL further NB-LRR types are known with additional N- or C-terminal domains, such as PK-NB-LRR, HD-NB-LRR and TIR-NB-LRR-WRKY (Meyers et al. 2003;Xue et al. 2012). The approach presented in this thesis would be able to identify these as well, as long as the fragments are long enough and paired-end sequencing or a novel approach such as PacBio RS is applied. Literature is however not able to answer the question for a possible fragment length that can be pulled out by a single bait.

Chapter 6 - Future perspectives

The results achieved in the studies presented in this thesis have answered some key questions about the abundance of NB-LRR genes in potato and their genomic organisation. Furthermore, a powerful NB-LRR gene enrichment and sequencing platform has been developed that has proven to be suitable for the identification of new NB-LRRs from a sequenced genome. It is envisaged, and some preliminary evidence already exists, that this platform will enable rapid mapping and cloning of functional NB-LRR-type resistance genes from diploid segregating populations. However, based on my research, the following questions have arisen and could be addressed in follow-on studies:

1) NB-LRR gene discovery and annotation in DM

- The NB-LRR gene enrichment and sequencing-based identification of novel DM NB-LRR genes yielded 338 potential genomic regions containing the coding sequences of these genes. For a wide range of further analyses including phylogenetic studies and the design of new bait libraries, it will be necessary to establish the exon sequences. Work that has been initiated to support newly and previously annotated NB-LRR gene models is the mapping of available RNA-seq data (PGSC 2011). The RNA-seq libraries available contain some derived from *Phytophthora infestans* infected plant material of DM and are currently being mapped against the genome superscaffolds. Changes in mapping stringencies will help identifying also the orthologues that are not expressed. Mapping information will yield exact coordinates of the coding regions. However, owing to the often reported low and constitutive expression of NB-LRR genes, where insufficient RNA-seq coverage is available, other approaches such as tblastx searches with manually curated genes could be utilised. Furthermore, MEME and MAST searches could be utilised to aid the NB-LRR gene annotation as described in Chapter 2.

- The above mentioned prediction of coding sequences for the novel NB-LRRs, followed by an amino acid guided nucleotide sequence alignment of the NB-ARC domain as carried out in chapter 2, will aid the identification of phylogenetic groups to which the new NB-LRR genes belong. Furthermore, alignments of the LRR domain and analysis of arising conserved groups will shed further light on to NB-LRR recognition specificities, and potentially identify groups with novel binding sites. This work will also shed light on

the diversification of NB-LRR genes both within and, ultimately, between Solanaceae species.

- Approximately 17% of the DM NB-LRR genes are located on yet unanchored superscaffolds. For a better understanding of NB-LRR gene clustering and their dispersal, final mapping of these unanchored sequences to the 12 DM chromosomes is necessary. Important information for this task might be retrieved from the ongoing potato genome re-mapping project (Dr Glenn Bryan personal communication).

- The number of NB-LRR encoding genes identified from the tomato genome sequence is less than a third of the number of potato NB-LRRs. Using the here presented NB-LRR capture approach on the sequenced tomato clone might, similar to DM, aid the identification of numerous NB-LRRs in poorly or incorrectly annotated genome regions. The high level of synteny reported for potato and tomato could also be exploited by simply mapping the DM Illumina reads obtained in this study to the tomato genome followed by an analysis of the read coverage over the already annotated tomato NB-LRRs. Similar to the described DM analysis, regions with high read depth could then be extracted and annotated using MAST and blast searches.

2) Further developing the NB-LRR gene enrichment procedure to map and clone functional NB-LRR genes

- Though the NB-LRR gene enrichment and Illumina sequencing proved very successful in the sequenced monoploid potato clone DM, the approach was not as successful in the tetraploid potato cultivar, Sarpo Mira. It appears that the short, 76bp reads made it impossible to distinguish between allelic and paralogous NB-LRR gene sequences in Sarpo Mira. Similar problems, though not to the same extent, are envisaged for diploid potato species. However, new sequencing technologies such as PacBio RS and Illumina MiSeq are emerging and longer sequencing read-length of up to 3kb and 250bp can be achieved, respectively (Quail et al. 2012). Therefore, it would be worth-while repeating the NB-LRR gene enrichment procedure for Sarpo Mira and Maris Piper to establish the full NB-LRR gene complement in these two cultivars. The resistant gene complements could then be used to map the less error-prone shorter reads from bulk resistant and bulk susceptible plants at high stringency to the parental NB-LRR gene templates to identify candidate NB-LRRs, and finally call polymorphisms between these.

- Cluster specific NB-LRR bait libraries could be developed to rapidly clone functional NB-LRRs if their genetic position coincides with well described genomic regions.

- As the '100 *Solanum'* genome sequencing project is advancing (http://solgenomics.net/organism/sol100/view), additional NB-LRR specific baits from these species could be included to generate a universal, *Solanum* applicable NB-LRR enrichment library.

3) Further characterisation of the resistance mechanisms of Sarpo Mira

- A recent paper by Rietman et al. (2012) identified the recognition of five *P. infestans* RXLR effectors in Sarpo Mira. Experiments could be carried out to determine a potential recognition of these and co-segregation with the resistance to the late blight isolates 3928A and 7454A seen in whole plant assays carried out in this thesis. Furthermore, expression analysis from various infection stages should be carried out, also to validate data by Cooke et al. (2012). Cooke et al. (2012) presented the identification of six novel RXLR effectors from isolate 3928A, absent from the range tested by Rietman and colleagues. Testing further the 6_A1 isolates for presence, followed by a transient expression of these six effectors in Sarpo Mira as well as the most resistant and susceptible plants of the SM×MP population (as defined in chapter 4) might identify a novel avirulence gene, mediating the resistance against 3928A and/or 7454A. The identification of RXLR effectors which, upon detection, yield resistance is an important tool to functionally verify candidate NB-LRR genes by co-expression in the model species *Nicotiana benthamiana* (Armstrong et al. 2005;Vleeshouwers et al. 2008).

- As a result of the Sarpo Mira × Maris Piper population genotyping, two loci were identified in the proximity of recently reported miRNAs that have been identified as suppressors of NB-LRR genes. It would be interesting to investigate if these miRNAs are differentially regulated and play a role in disease resistance/susceptibility. This could be studied by using for example northern blot experiments or miRNA specific arrays (e.g. from Affymetrix). A potential reduction of miRNA levels in Sarpo Mira would indicate an involvement of these in the resistance towards late blight.

- The phenotypic data derived from the screening of the Sarpo Mira × Maris Piper cross with the *P. infestans* isolate 7454A could be integrated with the Illumina

GoldenGate genotyping data. Knowledge of the underlying resistance loci will then potentially corroborate identified genetic positions for the resistance loci against *P. infestans* isolate 3928A.

- The Illumina sequence reads derived during chapter 4 for the NB-LRR enriched BS and BR can be re-analysed using a new bioinformatical pipeline for SNP-calling currently under development at The Sainsbury Laboratory (K. Witek personal communication). This pipeline uses a combination of mapping reads to the reference DM, as well as *de novo* assemblies of reads derived from the parents. To these assemblies BS and BR reads could be mapped to identify polymorphisms that are present at a ratio of approximately 25% in the resistant parent and BR, and less than 5% in the susceptible parent and BS. Verification of the derived polymorphisms would be carried out by amplification from single accessions of these bulks and Sanger sequencing. This analysis will help to confirm the genotyping results achieved in chapter 4, or the results achieved by Kare Lehman Nielsen that were briefly discussed.

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J.Mol.Biol.* 215, no. 3:403-410.

Anderson, L. K., A. Lai, S. M. Stack, C. Rizzon, and B. S. Gaut. 2006. Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome Res.* 16, no. 1:115-122.

Andolfo, G., W. Sanseverino, S. Rombauts, Y. Van de Peer, J. M. Bradeen, D. Carputo, L. Frusciante, and M. R. Ercolano. 2013. Overview of tomato (*Solanum lycopersicum*) candidate pathogen recognition genes reveals important *Solanum* R locus dynamics. *New Phytol.* 197, no. 1:223-237.

Armstrong, M. R., S. C. Whisson, L. Pritchard, J. I. Bos, E. Venter, A. O. Avrova, A. P. Rehmany et al. 2005. An ancestral oomycete locus contains late blight avirulence gene Avr3a, encoding a protein that is recognized in the host cytoplasm. *Proc.Natl.Acad.Sci.U.S.A* 102, no. 21:7766-7771.

Bailey, T. L., M. Boden, T. Whitington, and P. Machanick. 2010. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* 11:179.

Bailey, T. L. and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology,* pp. 28-36.

Bailey, T. L. and M. Gribskov. 1998. Methods and statistics for combining motif match scores. *J.Comput.Biol.* 5, no. 2:211-221.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One.* 3, no. 10:e3376.

Bakker, E., T. Borm, P. Prins, E. A. G. van der Vossen, G. Uenk, M. Arens, J. de Boer et al. 2011. A genome-wide genetic map of NB-LRR disease resistance loci in potato. *Theor.Appl.Genet.* 123, no. 3:493-508.

Ballvora, A., M. R. Ercolano, J. Weiss, K. Meksem, C. A. Bormann, P. Oberhagemann, F. Salamini, and C. Gebhardt. 2002. The R1 gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *Plant J.* 30, no. 3:361-371.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones et al. 2002. The Pfam protein families database. *Nucleic Acids Res.* 30, no. 1:276-280.

Baumgarten, A., S. Cannon, R. Spangler, and G. May. 2003. Genome-level evolution of resistance genes in *Arabidopsis thaliana*. *Genetics* 165, no. 1:309-319.

Belkhadir, Y., R. Subramaniam, and J. L. Dangl. 2004. Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Curr.Opin. Plant Biol.* 7, no. 4:391-399.

Bendahmane, A., M. Querci, K. Kanyuka, and D. C. Baulcombe. 2000. Agrobacterium transient expression system as a tool for the isolation of disease resistance genes: application to the *Rx2* locus in potato. *Plant J.* 21, no. 1:73-81.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, no. 7218:53-59.

Biesecker, L. G., K. V. Shianna, and J. C. Mullikin. 2011. Exome sequencing: the expert view. *Genome Biol.* 12, no. 9:128.

Blankenberg, D., A. Gordon, G. Von Kuster, N. Coraor, J. Taylor, and A. Nekrutenko. 2010a. Manipulation of FASTQ data with Galaxy. *Bioinformatics* 26, no. 14:1783-1785.

Blankenberg, D., G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. 2010b. Galaxy: a web-based genome analysis tool for experimentalists. *Curr.Protoc.Mol.Biol.* Chapter 19:Unit 19.10.1-21. doi: 10.1002/0471142727.mb1910s89

Boller, T. and G. Felix. 2009. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annu.Rev. Plant Biol.* 60:379-406.

Boller, T. and S. Y. He. 2009. Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* 324, no. 5928:742-744.

Bormann, C. A., A. M. Rickert, R. A. Ruiz, J. Paal, J. Lubeck, J. Strahwald, K. Buhr, and C. Gebhardt. 2004. Tagging quantitative trait loci for maturity-corrected late blight resistance in tetraploid potato with PCR-based candidate gene markers. *Mol. Plant Microbe Interact.* 17, no. 10:1126-1138.

Bos, J. I. B., M. R. Armstrong, E. M. Gilroy, P. C. Boevink, I. Hein, R. M. Taylor, Tian Z. D. et al. 2010. *Phytophthora infestans* effector AVR3a is essential for virulence and manipulates plant immunity by stabilizing host E3 ligase CMPG1. *Proc. Natl. Acad. Sci. U.S.A*. 107 21:9909-9914

Bradshaw, J. E., G. J. Bryan, A. K. Lees, K. McLean, and R. M. Solomon-Blackburn. 2006. Mapping the *R10* and *R11* genes for resistance to late blight (*Phytophthora infestans*) present in the potato (*Solanum tuberosum*) *R*-gene differentials of Black. *Theor.Appl.Genet.* 112, no. 4:744-751.

Bradshaw, J. E., B. Pande, G. J. Bryan, C. A. Hackett, K. McLean, H. E. Stewart, and R. Waugh. 2004. Interval mapping of quantitative trait loci for resistance to late blight [*Phytophthora infestans* (Mont.) de Bary], height and maturity in a tetraploid population of potato (*Solanum tuberosum* subsp. *tuberosum*). *Genetics* 168, no. 2:983-995.

Brommonschenkel, S. H., A. Frary, and S. D. Tanksley. 2000. The broad-spectrum tospovirus resistance gene *Sw-5* of tomato is a homolog of the root-knot nematode resistance gene *Mi*. *Mol.Plant Microbe Interact.* 13, no. 10:1130-1138.

Bryan, G. J. and I. Hein. 2008. Genomic resources and tools for gene function analysis in potato. *Int.J.Plant Genomics* 2008:216513.

Cannon, S. B., H. Zhu, A. M. Baumgarten, R. Spangler, G. May, D. R. Cook, and N. D. Young. 2002. Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. *J.Mol.Evol.* 54, no. 4:548-562.

Chen, Y., Z. Liu, and D. A. Halterman. 2012. Molecular determinants of resistance activation and suppression by *Phytophthora infestans* effector *IPI-O*. *PLoS Pathog.* 8, no. 3:e1002595.

Chinchilla, D., Z. Bauer, M. Regenass, T. Boller, and G. Felix. 2006. The Arabidopsis receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception. *Plant Cell* 18, no. 2:465-476.

Chini, A. and G. J. Loake. 2005. Motifs specific for the ADR1 NBS-LRR protein family in Arabidopsis are conserved among NBS-LRR sequences from both dicotyledonous and monocotyledonous plants. *Planta* 221, no. 4:597-601.

Chisholm, S. T., G. Coaker, B. Day, and B. J. Staskawicz. 2006. Host-microbe interactions: shaping the evolution of the plant immune response. *Cell* 124, no. 4:803-814.

Close, T. J., P. R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka et al. 2009. Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* 10:582.

Cock, P. J., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 25, no. 11:1422-1423.

Collier, S. M., L. P. Hamel, and P. Moffett. 2011. Cell death mediated by the N-terminal domains of a unique and highly conserved class of NB-LRR protein. *Mol.Plant Microbe Interact.* 24, no. 8:918-931.

Collier, S. M. and P. Moffett. 2009. NB-LRRs work a "bait and switch" on pathogens. *Trends Plant Sci.* 14, no. 10:521-529.

Colon, L., R. Solomon-Blackburn, B. J. Nielsen, and U. Darsow. 2004. Eucablight protocol - Whole plant glasshouse test for foliage blight resistance Version 1.2. 4-12-2012.

Colon, L. T., D. E. L. Cooke, J. G. Hansen, P. Lassen, D. Andrivon, A. Hermansen, E. Zimnoch-Guzowska, and A. K. Lees. 2005. Eucablight: a late blight network for Europe. In *Potato in progress. Science meets practice.*, eds. Haverkort, A. J. and P. C. Struik, 290-298. Wageningen Academic Publishers).

Cooke, D. E., L. M. Cano, S. Raffaele, R. A. Bain, L. R. Cooke, G. J. Etherington, K. L. Deahl et al. 2012. Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. *PLoS Pathog.* 8, no. 10:e1002940.

Cruickshank, G., H. E. Stewart, and R. L. Wastie. 1982. An illustrated assessment key for foliage blight of potatoes. *Potato Research* 25:213-214.

Cui, H., T. Xiang, and J. M. Zhou. 2009. Plant immunity: a lesson from pathogenic bacterial effector proteins. *Cell Microbiol.* 11, no. 10:1453-1461.

Dangl, J. L. and J. D. Jones. 2001. Plant pathogens and integrated defence responses to infection. *Nature* 411, no. 6839:826-833.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat.Rev.Genet.* 12, no. 7:499-510.

de Jonge, R., H. P. van Esse, A. Kombrink, T. Shinya, Y. Desaki, R. Bours, S. van der Krol et al. 2010. Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. *Science* 329, no. 5994:953-955.

de Jonge, R., H. P. van Esse, K. Maruthachalam, M. D. Bolton, P. Santhanam, M. K. Saber, Z. Zhang et al. 2012. Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proc.Natl.Acad.Sci.U.S.A* 109, no. 13:5110-5115.

Deslandes, L. and S. Rivas. 2012. Catch me if you can: bacterial effectors and plant targets. *Trends Plant Sci.* 17, no. 11:644-655.

DeYoung, B. J., D. Qi, S. H. Kim, T. P. Burke, and R. W. Innes. 2012. Activation of a plant nucleotide binding-leucine rich repeat disease resistance protein by a modified self protein. *Cell Microbiol.* 14, no. 7:1071-1084.

Eddy, S. R. 2008. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput.Biol.* 4, no. 5:e1000069.

el-Kharbotly, A., C. Leonards-Schippers, D. J. Huigen, E. Jacobsen, A. Pereira, W. J. Stiekema, F. Salamini, and C. Gebhardt. 1994. Segregation analysis and RFLP mapping of the *R1* and *R3* alleles conferring race-specific resistance to *Phytophthora infestans* in progeny of dihaploid potato parents. *Mol.Gen.Genet.* 242, no. 6:749-754.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One.* 6, no. 5:e19379.

Ernst, K., A. Kumar, D. Kriseleit, D. U. Kloos, M. S. Phillips, and M. W. Ganal. 2002. The broad-spectrum potato cyst nematode resistance gene (*Hero*) from tomato is the only member of a large gene family of NBS-LRR genes with an unusual amino acid repeat in the LRR region. *Plant J.* 31, no. 2:127-136.

Felix, G., J. D. Duran, S. Volko, and T. Boller. 1999. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *Plant J.* 18, no. 3:265-276.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38, no. Database issue:D211-D222.

Flor, H. H. 1942. Inheritance of pathogenicity in *Melampsora lini*. *Phytopath.* 32:653-669.

Foster, S. J., T. H. Park, M. Pel, G. Brigneti, J. Sliwka, L. Jagger, der van, V, and J. D. Jones. 2009. *Rpi-vnt1.1*, a *Tm-2*(2) homolog from *Solanum venturii*, confers resistance to potato late blight. *Mol.Plant Microbe Interact.* 22, no. 5:589-600.

Freeman B. C. and G. A.Beattie. 2008. An overview over plant defences against pathogens and herbivores. *The Plant Health Instructor*. DOI: 10.1094/PHI-I-2008-0226-01

Friedman, A. R. and B. J. Baker. 2007. The evolution of resistance genes in multi-protein plant resistance systems. *Curr.Opin.Genet.Dev.* 17, no. 6:493-499.

Fry, W. 2008. *Phytophthora infestans*: the plant (and *R* gene) destroyer. *Mol.Plant Pathol.* 9, no. 3:385-402.

Gassmann, W. and S. Bhattacharjee. 2012. Effector-triggered immunity signaling: from gene-for-gene pathways to protein-protein interaction networks. *Mol.Plant Microbe Interact.* 25, no. 7:862-868.

Gebhardt, C. and J. P. T. Valkonen. 2001. Organization of genes controlling disease resistance in the potato genome. *Annu.Rev.Phytopathol.* 39:79-102.

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

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* 15, no. 10:1451-1455.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat.Biotechnol.* 27, no. 2:182-189.

Goecks, J., A. Nekrutenko, and J. Taylor. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11, no. 8:R86.

Golas, T. M., A. Sikkema, J. Gros, R. M. C. Feron, R. G. van den Berg, G. M. van den Weerden, C. Mariani, and J. J. H. M Allefs. 2010. Identification of a resistance gene *Rpi-dlc1* to *Phytophthora infestans* in European accessions of *Solanum dulcamara*. *Theor.Appl.Genet.* 120:797-808.

Goodwin, S. B. and A. Drenth. 1997. Origin of the A2 Mating Type of *Phytophthora infestans* Outside Mexico. *Phytopathology* 87, no. 10:992-999.

Goverse, A. and P. C. Struik. 2009. BIOEXPLOIT: Exploitation of Natural Plant Diversity for the Pesticide-Free Production of Food. *Potato Research* 52:207-208.

Greenville-Briggs, L. J. and P. Van West. 2005. The biotrophic stages of oomycete-plant interactions. In *Advances in Applied Microbiology*, 217-243. Elsevier Academic Press.

Guo, Y. L., J. Fitz, K. Schneeberger, S. Ossowski, J. Cao, and D. Weigel. 2011. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*. *Plant Physiol.* 157, no. 2:757-769.

Gutierrez, J. R., A. L. Balmuth, V. Ntoukakis, T. S. Mucyn, S. Gimenez-Ibanez, A. M. Jones, and J. P. Rathjen. 2010. *Prf* immune complexes of tomato are oligomeric and

contain multiple Pto-like kinases that diversify effector recognition. *Plant J.* 61, no. 3:507-518.

Haas, B. J., S. Kamoun, M. C. Zody, R. H. Jiang, R. E. Handsaker, L. M. Cano, M. Grabherr et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461, no. 7262:393-398.

Hackett, C. A. and Z. W. Luo. 2003. TetraploidMap: construction of a linkage map in autotetraploid species. *J.Hered.* 94, no. 4:358-359.

Hackett, C. A., I. Milne, J. E. Bradshaw, and Z. Luo. 2007. TetraploidMap for Windows: linkage map construction and QTL mapping in autotetraploid species. *J.Hered.* 98, no. 7:727-729.

Hammond-Kosack, K. E., D. A. Jones, and J. Jones. 1994. Identification of Two Genes Required in Tomato for Full Cf-9-Dependent Resistance to *Cladosporium fulvum*. *Plant Cell* 6, no. 3:361-374.

Haverkort, A. J., P. M. Boonekamp, R. C. B. Hutten, E. Jacobsen, L. A. P. Lotz, G. J. T. Kessel, R. G. F. Visser, and E. A. G. van der Vossen. 2008. Societal Costs of Late Blight in Potato and Prospects of Durable Resistance Through Cisgenic Modification. *Potato Res.* 51, no. 1:47-57.

Haverkort, A. J., P. C. Struik, R. G. Visser, and E. Jacobsen. 2009. Applied Biotechnology to Combat Late Blight in Potato Caused by *Phytophthora infestans*. *Potato Res.* 52:249-264.

Hawkes, J. G. 1993. Origins of Cultivated Potatoes and Species Relationships. In *Potato Genetics*, eds. Bradshaw, J. E. and G. R. Mackay, 3-42. (Oxford: CAB International).

Hein, I., P. R. J. Birch, S. Danan, V. Lefebvre, D. Achieng Odeny, C. Gebhardt, F. Trognitz, and G. J. Bryan. 2009a. Progress in Mapping and Cloning Qualitative and Quantitative Resistance Against *Phytophthora infestans* in Potato and Its Wild Relatives. *Potato Res.* 52:215-227.

Hein, I., E. M. Gilroy, M. R. Armstrong, and P. R. J. Birch. 2009b. The zig-zag-zig in oomycete-plant interactions. *Mol.Plant Pathol.* 10, no. 4:547-562.

Hein, I., K. McLean, B. Chalhoub, and G. J. Bryan. 2007. Generation and screening of a BAC library from a diploid potato clone to unravel durable late blight resistance on linkage group IV. *Int.J.Plant Genomics* 2007:51421.

Hordijk, W. and O. Gascuel. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics* 21, no. 24:4338-4347.

Hormaza, J. I., L. Dollo, and V. S. Polito. 1994. Identification of a RAPD marker linked to sex determination in *Pistacia vera* using bulked segregant analysis. *Theor.Appl.Genet.* 89, no. 1:9-13.

Huang, S., E. A. G. van der Vossen, H. Kuang, V. G. Vleeshouwers, N. Zhang, T. J. Borm, H. J. van Eck et al. 2005. Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato. *Plant J.* 42, no. 2:251-261.

Huang, S., V. G. Vleeshouwers, J. S. Werij, R. C. Hutten, H. J. van Eck, R. G. Visser, and E. Jacobsen. 2004. The R3 resistance to *Phytophthora infestans* in potato is conferred by two closely linked R genes with distinct specificities. *Mol.Plant Microbe Interact.* 17, no. 4:428-435.

Hulbert, S. H., C. A. Webb, S. M. Smith, and Q. Sun. 2001. Resistance gene complexes: evolution and utilization. *Annu.Rev.Phytopathol.* 39:285-312.

Ishibashi, K., K. Masuda, S. Naito, T. Meshi, and M. Ishikawa. 2007. An inhibitor of viral RNA replication is encoded by a plant resistance gene. *Proc.Natl.Acad.Sci.U.S.A* 104, no. 34:13833-13838.

Jacobs, M. M. J, B. Vosman, V. G. Vleeshouwers, R. G. Visser, B. Henken, and R. G. van den Berg. 2010. A novel approach to locate *Phytophthora infestans* resistance genes on the potato genetic map. *Theor.Appl.Genet.* 120:785-796.

Jiang, S. M., J. Hu, W. B. Yin, Y. H. Chen, R. R. Wang, and Z. M. Hu. 2005. Cloning of resistance gene analogs located on the alien chromosome in an addition line of wheat-*Thinopyrum intermedium*. *Theor.Appl.Genet.* 111, no. 5:923-931.

Jo, K. R., M. Arens, T. Y. Kim, M. A. Jongsma, R. G. Visser, E. Jacobsen, and J. H. Vossen. 2011. Mapping of the *S. demissum* late blight resistance gene *R8* to a new locus on chromosome IX. *Theor.Appl.Genet.* 123, no. 8:1331-1340.

Jones, J. D. and J. L. Dangl. 2006. The plant immune system. *Nature* 444, no. 7117:323-329.

Joosten, M. H., T. J. Cozijnsen, and P. J. de Wit. 1994. Host resistance to a fungal tomato pathogen lost by a single base-pair change in an avirulence gene. *Nature* 367, no. 6461:384-386.

Jupe, F., L. Pritchard, G. J. Etherington, K. Mackenzie, P. J. Cock, F. Wright, S. K. Sharma et al. 2012. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC.Genomics* 13:75.

Kawchuk, L. M., J. Hachey, D. R. Lynch, F. Kulcsar, G. van Rooijen, D. R. Waterer, A. Robertson et al. 2001. Tomato *Ve* disease resistance genes encode cell surface-like receptors. *Proc.Natl.Acad.Sci.U.S.A* 98, no. 11:6511-6515.

Kay, S. and U. Bonas. 2009. How *Xanthomonas* type III effectors manipulate the host plant. *Curr.Opin.Microbiol.* 12, no. 1:37-43.

Kelley, B. S., S. J. Lee, C. M. Damasceno, S. Chakravarthy, B. D. Kim, G. B. Martin, and J. K. Rose. 2010. A secreted effector protein (SNE1) from *Phytophthora infestans* is a broadly acting suppressor of programmed cell death. *Plant J*. 62, no. 3:357-366

Kim, H. J., H. R. Lee, K. R. Jo, S. M. Mortazavian, D. J. Huigen, B. Evenhuis, G. Kessel et al. 2012. Broad spectrum late blight resistance in potato differential set plants *MaR8* and *MaR9* is conferred by multiple stacked *R* genes. *Theor.Appl.Genet.* 124, no. 5:923-935.

Krasileva, K. V., C. Zheng, L. Leonelli, S. Goritschnig, D. Dahlbeck, and B. J. Staskawicz. 2011. Global analysis of *Arabidopsis*/downy mildew interactions reveals prevalence of incomplete resistance and rapid evolution of pathogen recognition. *PLoS One* 6, no. 12:e28765.

Kuang, H., F. Wei, M. R. Marano, U. Wirtz, X. Wang, J. Liu, W. P. Shum et al. 2005. The *R1* resistance gene cluster contains three groups of independently evolving, type I *R1* homologues and shows substantial structural variation among haplotypes of *Solanum demissum*. *Plant J.* 44, no. 1:37-51.

Kuhl, J. C., R. E. Hanneman, Jr., and M. J. Havey. 2001. Characterization and mapping of *Rpi1*, a late-blight resistance locus from diploid (1EBN) Mexican *Solanum pinnatisectum*. *Mol.Genet Genomics* 265, no. 6:977-985.

Lanfermeijer, F. C., J. Dijkhuis, M. J. Sturre, P. de Haan, and J. Hille. 2003. Cloning and characterization of the durable tomato mosaic virus resistance gene Tm-2(2) from *Lycopersicon esculentum*. *Plant Mol.Biol.* 52, no. 5:1037-1049.

Langmead, B. 2010. Aligning short sequencing reads with Bowtie. *Curr.Protoc.Bioinformatics.* 32:11.7.1-11.7.14. John Wiley & Sons, Inc.

Leister, D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet.* 20, no. 3:116-122.

Li, F., D. Pignatta, C. Bendix, J. O. Brunkard, M. M. Cohn, J. Tung, H. Sun, P. Kumar, and B. Baker. 2012. MicroRNA regulation of plant innate immune receptors. *Proc.Natl.Acad.Sci.U.S.A* 109, no. 5:1790-1795.

Li, G., S. Huang, X. Guo, Y. Li, Y. Yang, Z. Guo, H. Kuang et al. 2011. Cloning and characterization of r3b; members of the r3 superfamily of late blight resistance genes show sequence and functional divergence. *Mol.Plant Microbe Interact.* 24, no. 10:1132-1142.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, no. 16:2078-2079.

Li, X., H. J. van Eck, J. N. A. M. Rouppe van der Voort, D. J. Huigen, P. Stam, and E. Jacobsen. 1998. Autotetraploids and genetic mapping using common AFLP markers: the *R2* allele conferring resistance to *Phytophthora infestans* mapped on potato chromosome 4. *Theor.Appl.Genet.* 96:1121-1128.

Liu, J., J. M. Elmore, Z. J. Lin, and G. Coaker. 2011. A receptor-like cytoplasmic kinase phosphorylates the host target RIN4, leading to the activation of a plant innate immune receptor. *Cell Host.Microbe* 9, no. 2:137-146.

Lokossou, A. A., T. H. Park, G. van Arkel, M. Arens, C. Ruyter-Spira, J. Morales, S. C. Whisson et al. 2009. Exploiting knowledge of R/Avr genes to rapidly clone a new LZ-NBS-LRR family of late blight resistance genes from potato linkage group IV. *Mol Plant Microbe Interact.* 22, no. 6:630-641.

Lozano, R., O. Ponce, M. Ramirez, N. Mostajo, and G. Orjeda. 2012. Genome-wide identification and mapping of NBS-encoding resistance genes in *Solanum tuberosum* group Phureja. *PLoS One* 7, no. 4:e34775.

Lukasik, E. and F. L. Takken. 2009. STANDing strong, resistance proteins instigators of plant defence. *Curr Opin.Plant Biol.* 12, no. 4:427-436.

Mackey, D., B. F. Holt, III, A. Wiig, and J. L. Dangl. 2002. RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. *Cell* 108, no. 6:743-754.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner. 2010. Target-enrichment strategies for next-generation sequencing. *Nat.Methods* 7, no. 2:111-118.

McDonnell, A. V., T. Jiang, A. E. Keating, and B. Berger. 2006. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics.* 22, no. 3:356-358.

McDowell, J. M. and S. A. Simon. 2006. Recent insights into R gene evolution. *Mol.Plant Pathol.* 7, no. 5:437-448.

McHale, L., X. Tan, P. Koehl, and R. W. Michelmore. 2006. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* 7, no. 4:212.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat.Rev.Genet.* 11, no. 1:31-46.

Meyers, B. C., A. W. Dickerman, R. W. Michelmore, S. Sivaramakrishnan, B. W. Sobral, and N. D. Young. 1999. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.* 20, no. 3:317-332.

Meyers, B. C., A. Kozik, A. Griego, H. Kuang, and R. W. Michelmore. 2003. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 15, no. 4:809-834.

Meyers, B. C., M. Morgante, and R. W. Michelmore. 2002. TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in *Arabidopsis* and other plant genomes. *Plant J.* 32, no. 1:77-92.

Michelmore, R. W. and B. C. Meyers. 1998. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* 8, no. 11:1113-1130.

Michelmore, R. W., I. Paran, and R. V. Kesseli. 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. *Proc.Natl.Acad.Sci.U.S.A* 88, no. 21:9828-9832.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, and E. A. Johnson. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 17, no. 2:240-248.

Milligan, S. B., J. Bodeau, J. Yaghoobi, I. Kaloshian, P. Zabel, and V. M. Williamson. 1998. The root knot nematode resistance gene *Mi* from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. *Plant Cell* 10, no. 8:1307-1319.

Milne, I., D. Lindner, M. Bayer, D. Husmeier, G. McGuire, D. F. Marshall, and F. Wright. 2009. TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics* 25, no. 1:126-127.

Mokry, M., I. J. Nijman, Dijken A. van, R. Benjamins, R. Heidstra, B. Scheres, and E. Cuppen. 2011. Identification of factors required for meristem function in *Arabidopsis* using a novel next generation sequencing fast forward genetics approach. *BMC Genomics* 12:256.

Mun, J. H., H. J. Yu, S. Park, and B. S. Park. 2009. Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa*. *Mol.Genet. Genomics* 282, no. 6:617-631.

Naess, S. K., J. M. Bradeen, S. M. Wielgus, G. T. Haberlach, J. M. McGrath, and J. P. Helgeson. 2000. Resistance to late blight in *Solanum bulbocastanum* is mapped to chromosme 8. *Theor.Appl.Genet.* 101:697-704.

Noutoshi, Y., T. Ito, M. Seki, H. Nakashita, S. Yoshida, Y. Marco, K. Shirasu, and K. Shinozaki. 2005. A single amino acid insertion in the WRKY domain of the *Arabidopsis* TIR-NBS-LRR-WRKY-type disease resistance protein SLH1 (sensitive to low humidity 1) causes activation of defense responses and hypersensitive cell death. *Plant J.* 43, no. 6:873-888.

Oosumi, T., D. R. Rockhold, M. M. Maccree, K. L. Deahl, K. F. McCue, and W. R. Belknap. 2009. Gene *Rpi-bt1* from *Solanum bulbocastanum* Confers Resistance to Late Blight in Transgenic Potatoes. *Americ.J.Potato Res.* 86, no. 6.

Ori, N., Y. Eshed, I. Paran, G. Presting, D. Aviv, S. Tanksley, D. Zamir, and R. Fluhr. 1997. The I2C family from the wilt disease resistance locus I2 belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. *Plant Cell* 9, no. 4:521-532.

Orlowska, E., A. Fiil, H. G. Kirk, B. Llorente, and C. Cvitanich. 2012. Differential gene induction in resistant and susceptible potato cultivars at early stages of infection by *Phytophthora infestans*. *Plant Cell Rep.* 31, no. 1:187-203.

Paape, T., P. Zhou, A. Branca, R. Briskine, N. Young, and P. Tiffin. 2012. Fine-scale population recombination rates, hotspots, and correlates of recombination in the *Medicago truncatula* genome. *Genome Biol.Evol.* 4, no. 5:726-737.

Padmanabhan, M., P. Cournoyer, and S. P. Dinesh-Kumar. 2009. The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. *Cell Microbiol.* 11, no. 2:191-198.

Park, T. H., J. Gros, A. Sikkema, V. G. Vleeshouwers, M. Muskens, S. Allefs, E. Jacobsen, R. G. Visser, and E. A. G. van der Vossen. 2005. The late blight resistance locus *Rpi-blb3* from *Solanum bulbocastanum* belongs to a major late blight *R* gene cluster on chromosome 4 of potato. *Mol. Plant Microbe Interact.* 18, no. 7:722-729.

Parla, J. S., I. Iossifov, I. Grabill, M. S. Spector, M. Kramer, and W. R. McCombie. 2011. A comparative analysis of exome capture. *Genome Biol.* 12, no. 9:R97.

Parniske, M., K. E. Hammond-Kosack, C. Golstein, C. M. Thomas, D. A. Jones, K. Harrison, B. B. Wulff, and J. D. Jones. 1997. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell* 91, no. 6:821-832.

Peart, J. R., P. Mestre, R. Lu, I. Malcuit, and D. C. Baulcombe. 2005. NRG1, a CC-NB-LRR protein, together with N, a TIR-NB-LRR protein, mediates resistance against tobacco mosaic virus. *Curr.Biol.* 15, no. 10:968-973.

Pel, M. A., S. J. Foster, T. H. Park, H. Rietman, G. van Arkel, J. D. Jones, H. J. van Eck et al. 2009. Mapping and cloning of late blight resistance genes from *Solanum venturii* using an interspecific candidate gene approach. *Mol.Plant Microbe Interact.* 22, no. 5:601-615.

PGSC The Potato Genome Sequencing Consortium 2011. Genome sequence and analysis of the tuber crop potato. *Nature* 475, no. 7355:189-195.

Postma, W. J., E. J. Slootweg, S. Rehman, A. Finkers-Tomczak, T. O. Tytgat, K. van Gelderen, J. L. Lozano-Torres et al. 2012. The effector SPRYSEC-19 of *Globodera rostochiensis* suppresses CC-NB-LRR-mediated disease resistance in plants. *Plant Physiol.* 160, no. 2:944-954.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341.

Quarrie, S. A., V. Lazic-Jancic, D. Kovacevic, A. Steed, and S. Pekic. 1999. Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. *J.Exp.Bot.* 50, no. 337:1299-1306.

Raffaele, S. and S. Kamoun. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat.Rev.Microbiol.* 10, no. 6:417-430.

Rairdan, G. J., S. M. Collier, M. A. Sacco, T. T. Baldwin, T. Boettrich, and P. Moffett. 2008. The coiled-coil and nucleotide binding domains of the Potato Rx disease resistance protein function in pathogen recognition and signaling. *Plant Cell* 20, no. 3:739-751.

Rauscher, G. M., C. D. Smart, I. Simko, M. Bonierbale, H. Mayton, A. Greenland, and W. E. Fry. 2006. Characterization and mapping of *Rpi-ber*, a novel potato late blight resistance gene from *Solanum berthaultii*. *Theor.Appl.Genet.* 112, no. 4:674-687.

Rietman, H., G. Bijsterbosch, L. M. Cano, H. R. Lee, J. H. Vossen, E. Jacobsen, R. G. Visser, S. Kamoun, and V. G. Vleeshouwers. 2012. Qualitative and quantitative late blight resistance in the potato cultivar Sarpo Mira is determined by the perception of five distinct RXLR effectors. *Mol.Plant Microbe Interact.* 25, no. 7:910-919.

Rivas, S. and C. M. Thomas. 2005. Molecular interactions between tomato and the leaf mold pathogen *Cladosporium fulvum*. *Annu.Rev.Phytopathol.* 43:395-436.

Rodriguez, F., M. Ghislain, A. M. Clausen, S. H. Jansky, and D. M. Spooner. 2010. Hybrid origins of cultivated potato. *Theor.Appl.Genet.* 121:1187-1198.

Rossman, A. Y. and M. E. Palm. 2006. Why are Phytophthora and other Oomycota not true Fungi? In *Outlooks on Pest Management* 17:217-219.

Saintenac, C., S. Faure, A. Remay, F. Choulet, C. Ravel, E. Paux, F. Balfourier, C. Feuillet, and P. Sourdille. 2011. Variation in crossover rates across a 3-Mb contig of bread

wheat (*Triticum aestivum*) reveals the presence of a meiotic recombination hotspot. *Chromosoma* 120, no. 2:185-198.

Salaman, R. N. 1911. The inheritance of colour and other characters in the potato. *J.Genetics* 1:7-43.

Sarvari Trust 2012. www.sarvari-trust.org, accessed 10-4-2012.

Schornack, S., A. Ballvora, D. Gurlebeck, J. Peart, D. Baulcombe, M. Ganal, B. Baker, U. Bonas, and T. Lahaye. 2004. The tomato resistance protein Bs4 is a predicted non-nuclear TIR-NB-LRR protein that mediates defense responses to severely truncated derivatives of AvrBs4 and overexpressed AvrBs3. *Plant J.* 37, no. 1:46-60.

Schornack, S., E. Huitema, L. M. Cano, T. O. Bozkurt, R. Oliva, M. Van Damme, S. Schwizer et al. 2009. Ten things to know about oomycete effectors. *Mol.Plant Pathol.* 10, no. 6:795-803.

Schornack, S., M. Van Damme, T. O. Bozkurt, L. M. Cano, M. Smoker, M. Thines, E. Gaulin, S. Kamoun, and E. Huitema. 2010. Ancient class of translocated oomycete effectors targets the host nucleus. *Proc.Natl.Acad.Sci.U.S.A* 107, no. 40:17421-17426.

Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc.Natl.Acad.Sci.U.S.A* 95, no. 11:5857-5864.

Schulze-Lefert, P. and R. Panstruga. 2011. A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. *Trends Plant Sci.* 16, no. 3:117-125.

Schumann, G L and C J D'Arcy. The Plant Health Instructor. 2000. APSnet, DOI: 10.1094/PHI-I-2000-0724-01

Shattock, R. C. 2002. *Phytophthora infestans*: populations, pathogenicity and phenylamides. *Pest.Manag.Sci.* 58, no. 9:944-950.

Shaw, D. and L. Johnson. Progress in the selection of cultivars with resistance to late-blight disease. PPO-Special Report No 10, 203-209. 2004.

Shen, K. A., D. B. Chin, R. Arroyo-Garcia, O. E. Ochoa, D. O. Lavelle, T. Wroblewski, B. C. Meyers, and R. W. Michelmore. 2002. *Dm3* is one member of a large constitutively expressed family of nucleotide binding site-leucine-rich repeat encoding genes. *Mol. Plant Microbe Interact.* 15, no. 3:251-261.

Shivaprasad, P. V., H. M. Chen, K. Patel, D. M. Bond, B. A. Santos, and D. C. Baulcombe. 2012. A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. *Plant Cell* 24, no. 3:859-874.

Sliwka, J. 2004. Genetic factors encoding resistance to late blight caused by *Phytophthora infestans* (Mont.) de Bary on the potato genetic map. *Cell Mol.Biol.Lett.* 9, no. 4B:855-867.

Smilde, W. D., G. Brigneti, L. Jagger, S. Perkins, and J. D. Jones. 2005. *Solanum mochiquense* chromosome IX carries a novel late blight resistance gene *Rpi-moc1*. *Theor.Appl.Genet.* 110, no. 2:252-258.

Smith, K. R., C. J. Bromhead, M. S. Hildebrand, A. E. Shearer, P. J. Lockhart, H. Najmabadi, R. J. Leventer et al. 2011. Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol.* 12, no. 9:R85.

Smith, L. P. 1956. Potato blight forecasting by 90 per cent humidity criteria. *Plant Path.* 5, 83-87.

Song, J., J. M. Bradeen, S. K. Naess, J. A. Raasch, S. M. Wielgus, G. T. Haberlach, J. Liu et al. 2003. Gene RB cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight. *Proc.Natl.Acad.Sci.U.S.A* 100, no. 16:9128-9133.

Spooner, D. M. and A. M. Clausen. 1993. Wild potato (Solanum sect. Petota) germplasm collecting expedition to Argentina in 1990, and status of Argentinian potato germplasm resources. *Potato Res.* 36:3-12.

Stitziel, N. O., A. Kiezun, and S. Sunyaev. 2011. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* 12, no. 9:227.

Stower, H. 2011. The exome factor. *Genome Biol.* 12, no. 9:407.

Sulonen, A. M., P. Ellonen, H. Almusa, M. Lepisto, S. Eldfors, S. Hannula, T. Miettinen et al. 2011. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* 12, no. 9:R94.

Tai, T. H., D. Dahlbeck, E. T. Clark, P. Gajiwala, R. Pasion, M. C. Whalen, R. E. Stall, and B. J. Staskawicz. 1999. Expression of the Bs2 pepper gene confers resistance to bacterial spot disease in tomato. *Proc.Natl.Acad.Sci.U.S.A* 96, no. 24:14153-14158.

Takken, F. L. and W. I. Tameling. 2009. To nibble at plant resistance proteins. *Science* 324, no. 5928:744-746.

Tameling, W. I. and F. L. Takken. 2008. Resistance proteins: scouts of the plant innate immune system. *Eur.J. Plant Pathol.* 121:243-245.

Tan, M. Y., R. C. Hutten, C. Celis, T. H. Park, R. E. Niks, R. G. Visser, and H. J. van Eck. 2008. The R(Pi-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. *Mol. Plant Microbe Interact.* 21, no. 7:909-918.

Tan, M. Y., R. C. Hutten, R. G. Visser, and H. J. van Eck. 2010. The effect of pyramiding *Phytophthora infestans* resistance genes *Rpi-mcd1* and *Rpi-ber* in potato. *Theor.Appl.Genet.* 121, no. 1:117-125.

Tarr, D. E. and H. M. Alexander. 2009. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC Res.Notes* 2:197.

TGC Tomato Genome Consortium 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, no. 7400:635-641.

Thomma, B. P., T. Nurnberger, and M. H. Joosten. 2011. Of PAMPs and effectors: the blurred PTI-ETI dichotomy. *Plant Cell* 23, no. 1:4-15.

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

Trognitz, F. C. and B. R. Trognitz. 2005. Survey of resistance gene analogs in *Solanum caripense*, a relative of potato and tomato, and update on *R* gene genealogy. *Mol.Genet. Genomics* 274, no. 6:595-605.

Turkensteen, L. J. 1993. Durable resistance of potatoes against *Phytophthora infestans*. In *Durability of disease resistance*, eds. Jacobs, Th. and J. E. Parlevliet, 115-124. Kluwer Academic Publisher.

Turkensteen, L. J., W. G. Flier, R. Wanningen, and A. Mulder. 2000. Production, survival and infectivity of oospores of *Phytophthora infestans*. *Plant Pathology* 49, no. 6:688-696.

Umareus, V and M. Umareus. 1994. Inheritance of Resistance to Late Blight. In *Potato Genetics*, eds. Bradshaw, J. E. and G. R. Mackay, 365-402. CABI.

van den Burg, H. A., S. J. Harrison, M. H. Joosten, J. Vervoort, and P. J. de Wit. 2006. *Cladosporium fulvum* Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. *Mol.Plant Microbe Interact.* 19, no. 12:1420-1430.

van der Biezen, E. A. and J. D. Jones. 1998a. Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem.Sci.* 23, no. 12:454-456.

van der Biezen, E. A. and J. D. Jones. 1998b. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr.Biol.* 8, no. 7:R226-R227.

van der Hoorn, R. A., F. Laurent, R. Roth, and P. J. de Wit. 2000. Agroinfiltration is a versatile tool that facilitates comparative analyses of Avr9/Cf-9-induced and Avr4/Cf-4-induced necrosis. *Mol.Plant Microbe Interact.* 13, no. 4:439-446.

van der Linden, C. G., D. C. Wouters, V. Mihalka, E. Z. Kochieva, M. J. Smulders, and B. Vosman. 2004. Efficient targeting of plant disease resistance loci using NBS profiling. *Theor.Appl.Genet.* 109, no. 2:384-393.

van der Vossen, E., J. Gros, A. Sikkema, M. Muskens, D. Wouters, P. Wolters, A. Pereira, and S. Allefs. 2005. The Rpi-blb2 gene from *Solanum bulbocastanum* is a *Mi-1* gene homolog conferring broad-spectrum late blight resistance in potato. *Plant J.* 44, no. 2:208-222.

van der Vossen, E., A. Sikkema, B. L. Hekkert, J. Gros, P. Stevens, M. Muskens, D. Wouters et al. 2003. An ancient R gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to Phytophthora infestans in cultivated potato and tomato. *Plant J.* 36, no. 6:867-882.

van Poppel, P. M., R. H. Jiang, J. Sliwka, and F. Govers. 2009. Recognition of *Phytophthora infestans* Avr4 by potato R4 is triggered by C-terminal domains comprising W motifs. *Mol.Plant Pathol.* 10, no. 5:611-620.

VanEtten, H. D., J. W. Mansfield, J. A. Bailey, and E. E. Farmer. 1994. Two Classes of Plant Antibiotics: Phytoalexins versus "Phytoanticipins". *Plant Cell* 6, no. 9:1191-1192.

Verzaux, E., D. Budding, N. Vetten, R. E. Niks, V. G. A. A. Vleeshouwers, E. A. G. van der Vossen, E. Jacobsen, and R. G. F. Visser. 2011. High Resolution Mapping of a Novel Late Blight Resistance Gene *Rpi-avl1*, from the Wild Bolivian Species *Solanum avilesii*. *Amer.J.Potato Res.* 88, no. 6:511-519.

Vleeshouwers, V. G., S. Raffaele, J. H. Vossen, N. Champouret, R. Oliva, M. E. Segretin, H. Rietman et al. 2011. Understanding and exploiting late blight resistance in the age of effectors. *Annu.Rev.Phytopathol.* 49:507-531.

Vleeshouwers, V. G., H. Rietman, P. Krenek, N. Champouret, C. Young, S. K. Oh, M. Wang et al. 2008. Effector genomics accelerates discovery and functional profiling of potato disease resistance and *Phytophthora infestans* avirulence genes. *PLoS ONE* 3, no. 8:e2875.

Vleeshouwers, V. G., W. van Dooijeweert, L. C. P. Keizer, L. Sijpkes, F. Govers, and L. T. Colon. 1999. A laboratory assay for *Phytophthora infestans* resistance in various *Solanum* species reflects the field situation. E*urop.J. Plant Pathol.* 105:241-250.

Wang, Q., C. Han, A. O. Ferreira, X. Yu, W. Ye, S. Tripathy, S. D. Kale et al. 2011. Transcriptional programming and functional interactions within the *Phytophthora sojae* RXLR effector repertoire. *Plant Cell* 23, no. 6:2064-2086.

Whisson, S. C., P. C. Boevink, L. Moleleki, A. O. Avrova, J. G. Morales, E. M. Gilroy, M. R. Armstrong et al. 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* 450, no. 7166:115-118.

White, S. and D. Shaw. Resistance of Sarpo clones to the new strains of *Phytophthora infestans*, Blue-13. PPO-Special Report No 13. 2009.

White, S. and D. Shaw. Breeding for host resistance: the key to sustainable potato production. PPO-Special Report 14, 125-132. 2010.

Xue, J. Y., Y. Wang, P. Wu, Q. Wang, L. T. Yang, X. H. Pan, B. Wang, and J. Q. Chen. 2012. A primary survey on bryophyte species reveals two novel classes of nucleotide-binding site (NBS) genes. *PLoS One.* 7, no. 5:e36700 .

Yang, S., T. Gu, C. Pan, Z. Feng, J. Ding, Y. Hang, J. Q. Chen, and D. Tian. 2008a. Genetic variation of NBS-LRR class resistance genes in rice lines. *Theor.Appl.Genet.* 116, no. 2:165-177.

Yang, S., X. Zhang, J. X. Yue, D. Tian, and J. Q. Chen. 2008b. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol.Genet. Genomics* 280, no. 3:187-198.

Zerbino, D. R. and E. Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, no. 5:821-829.

Zhang, G. Y., M. Chen, J. M. Guo, T. W. Xu, L. C. Li, Z. S. Xu, Y. Z. Ma, and X. P. Chen. 2009. Isolation and characteristics of the *CN* gene, a tobacco mosaic virus resistance *N* gene homolog, from tobacco. *Biochem.Genet.* 47, no. 3-4:301-314.

Zipfel, C., S. Robatzek, L. Navarro, E. J. Oakeley, J. D. Jones, G. Felix, and T. Boller. 2004. Bacterial disease resistance in *Arabidopsis* through flagellin perception. *Nature* 428, no. 6984:764-767.

Appendix

Annex 1 Trainingsets - Thesis Florian Jupe

Accession	Description	Type
	Negative training set	
CAA73187.1	Cf-4A {Solanum lycopersicum}	LRR
AAA65235.1	Cf-9 {Solanum lycopersicum}	LRR
AAX19019.1	Cf-2 {Solanum pimpinellifolium}	LRR
AAC78591.1	Cf-5 {Solanum lycopersicum}	LRR
CAL69642.1	RanGAP2 {Solanum tuberosum}	LRR
AAK58682.1	Ve1 {Solanum lycopersicum}	LRR
Q8L899.1	SR160 {Solanum peruvianum}	LRR
ABQ51814.1	RAN GTPase-activating protein 1	LRR
ABG74351.1	Ethylene-inducing xylanase {Solanum lycopersicum}	LRR
AAU90337.1	Putative receptor kinase-like protein {Solanum demissum}	LRR
AAT39281.2	Integrase core domain containing protein {Solanum demissum}	NB-ARC
AAT39966.1	Putative isopenicillin N epimerase {Solanum demissum}	NB-ARC
BAE06227.1	Heat shock protein {Solanum lycopersicum}	CC; ATPase
AAG39059.1	NBS-kinase protein Z2 {Solanum tuberosum}	NB-ARC
AAF74980.1	Cystathionine beta-lyase {Solanum tuberosum}	Other
AAY97864.1	Xyloglucan-specific endoglucanase inhibitor protein precursor {S. lycopersicum}	Other
	Positive training set	
AF121435.1	A5 {Capsicum annuum}	NB-ARC
AAF09256	BS2 {Capsicum chacoense}	NB-LRR
AAR21295	Bs4 {Solanum lycopersicum}	TIR-NB-LRR
CAB55838	Gpa2 {Solanum tuberosum subsp. andigenum}	NB-LRR
AAP44390	Gro1-4 {Solanum tuberosum}	TIR-NB-LRR
CAD29728	Hero {Solanum lycopersicum}	NB-LRR
AAD27815	I2 {Solanum lycopersicum}	CC-NB-LRR
AAC67238	Mi1.2 {Solanum lycopersicum}	CC-NB-LRR
ABO21407	N {Nicotiana tabacum}	TIR-NB-LRR
CAA08797	NL25 {Solanum tuberosum}	TIR-NB-LRR
CAA08798	NL27 {Solanum tuberosum}	TIR-NB-LRR
ABC26878	NRC1 {Solanum lycopersicum}	CC-NB-LRR
ABY61745	PSH-RGH6 {Solanum tuberosum}	NB-LRR
AAL39063	R1 {Solanum demissum}	NB-LRR
ACU65456	R2 {Solanum demissum}	CC-NB-LRR
AAW48299	R3a {Solanum demissum}	CC-NB-LRR
ACY74346	RGA2 {Capsicum annuum}	NB-LRR
AY336128	RGA2 {Solanum bulbocastanum}	NB-LRR
NP_199338.1	RPS4 {Arabidopsis thaliana}	TIR-NB-LRR
ACI16480	Rpi-bt1 {Solanum bulbocastanum}	NB-LRR
AAZ95005	Rpi-blb2 {Solanum bulbocastanum}	CC-NB-LRR
ACU65457.1	Rpi-blb3 {Solanum bulbocastanum}	CC-NB-LRR
ACJ66596	Rpi-vnt1.3 {Solanum venturii}	CC-NB-LRR
CAB50786	Rx {Solanum tuberosum}	NB-LRR
AAG31013	TVR-A {Solanum lycopersicum}	CC-NB-LRR
ABM05492	Tm-2 {Solanum tuberosum}	NB-LRR
CAC82811	Y-1 {Solanum tuberosum subsp. andigenum}	TIR-NB-LRR
AAM94164.1	Go35 NBS-LRR {Aegilops tauschii}	NB-LRR
AAF36987.1	Viral resistance protein {Arabidopsis thaliana}	NB-LRR

AAQ01784.1	Resistance protein LR10 {Triticum aestivum}	NB-LRR
ACZ65507.1	MLA1 {Hordeum chilense}	NB-LRR
AAQ55541.1	MLA10 {Hordeum vulgare}	CC-NB-LRR
AAO43441.1	MLA12 {Hordeum vulgare subsp. vulgare}	CC-NB-LRR
CAC29242.1	MLA6 protein {Hordeum vulgare subsp. vulgare}	CC-NB-LRR
ABC94599.1	NBS-LRR type R protein, Nbs4-Pi {Oryza sativa Indica Group}	NB-LRR
ABI64281.1	CC-NBS-LRR Pi36 {Oryza sativa Indica Group}	CC-NB-LRR
BAA76282.2	Pib {Oryza sativa Japonica Group}	NB-LRR
AAT08955.1	CC-NBS-LRR {Helianthus annuus}	CC-NB-LRR
AAQ96158.1	Powdery mildew resistance protein PM3b {Triticum aestivum}	CC-NB-LRR
AAD47197.1	Rust resistance protein {Zea mays}	NB-LRR
NP_187360.1	RPM1 (RESISTANCE TO P. SYRINGAE PV MACULICOLA 1){Arabidopsis thaliana}	NB-LRR
NP_190237.1	RPP13 (RECOGNITION OF PERONOSPORA PARASITICA 13) {Arabidopsis thaliana}	NB-LRR
NP_199160.1	RPP8 (RECOGNITION OF PERONOSPORA PARASITICA 8) {Arabidopsis thaliana}	CC-NB-LRR
AAX89382.1	NB-LRR type disease resistance protein Rps1-k-1 {Glycine max}	NB-LRR
AAX89383.1	NB-LRR type disease resistance protein Rps1-k-2 {Glycine max}	NB-LRR
NP_194339.1	RPS2 (RESISTANT TO P. SYRINGAE 2){Arabidopsis thaliana}	NB-LRR
NP_172686.1	RPS5 (RESISTANT TO P. SYRINGAE 5) {Arabidopsis thaliana}	NB-LRR
AAB47618.1	Rust resistance protein M {Linum usitatissimum}	TIR-NB-LRR
AAK28805.1	Resistance-like protein P2-A {Linum usitatissimum}	TIR-NB-LRR
NP_190034.2	RPP1 (recognition of peronospora parasitica 1) {Arabidopsis thaliana}	TIR-NB-LRR
NP_199338.1	RPS4 (RESISTANT TO P. SYRINGAE 4) {Arabidopsis thaliana}	TIR-NB-LRR
AAN86124.1	TIR-NBS-LRR {Arabidopsis thaliana}	TIR-NB-LRR
NP_001078715	RRS1 (RESISTANT TO RALSTONIA SOLANACEARUM 1){Arabidopsis thaliana}	TIR-NB-LRR

Annex 2 Protocol for NB-LRR capture enrichment experiment
Thesis Florian Jupe
This protocol is a composition of contents from manufacturer's protocols of Agilent Technologies and Agencourt, as well as Quail et al. (2009).

Pre hybridisation

Preparation of bulked DNA fragments

Shearing samples to ~500bp

For solution capture, dilute 3ug gDNA to a total volume of 100ul with water. If necessary, do Ethanol precipitation prior shearing.

1. Mix and transfer to a Covaris AFA fibre vial.
2. Seal the tube with meal crimp seal cap and crimping tool
3. Shear with Covaris, using the settings

Duty Cycle	20%
Intensity	5
Cycle/burst	200
Time	30sec
Temperature	4°C

4. Remove the sample from the machine. Open the vial and transfer the sample into a fresh 1.5ml Eppendorf lo-bind tube. Keep on Ice.

5. Run 1ul on 1.5% agarose gel to check size and on the Nanodrop to check quantity

Purification using SPRI beads

1. Let the SPRI beads come to room temperature

2. Mix reagents well!

3. Add 180 µL of homogenous SPRI XP beads to a 1.5-mL LoBind tube, and add the sheared DNA library (~120 µL). Mix well on a vortex mixer and incubate for 5 minutes.

4. Put the tube in the magnetic stand. Wait for the solution to clear (approximately 3 to 5 minutes).

5. Keep the tube in the magnetic stand. Do not touch the beads while you carefully discard the cleared solution from the tubes.

6. Continue to keep the tube in the magnetic stand while you dispense 500 µL of 70% ethanol in each tube.

7. Let the tube sit for 1 minute to allow any disturbed beads to settle, and remove the ethanol. Repeat this step once.

8. Dry the samples on the 37°C heat block for 5 minutes or until the residual ethanol completely evaporates.

9. Add 30 µL nuclease-free water, mix well on a vortex mixer, and incubate for 2 minutes at room temperature.

10. Put the tube in the magnetic stand and leave for 2 to 3 minutes, until the solution is clear.

11. Remove approximately 30 µL of the supernatant to a fresh 1.5-mL LoBind tube. You can discard the beads at this time.

Stopping Point If you do not continue to the next step, store the samples at -20°C.

End Repair of fragments

1. Prepare a reaction mix

DNA sample	30 µl	*-adjust*
Water	45 µl	*-adjust*
10x T4 DNA ligase buffer with 10mM ATP	10 µl	
10mM dNTP mix	4 µl	
3U/ µl T4 DNA polymerase	5 µl	
5U/ µl Klenow DNA polymerase	1 µl	
10U/ µl T4 PNK	5 µl	

2. Incubate for 30 minutes at room temperature (20°C).

3. Clean up using SPRI beads (180µl beads, as above). Elute in 31µl Sigma-water.

Addition of 'A' Base to the 3' End of the DNA Fragments

1. Prepare a reaction mix

DNA sample	30 µl
10x Klenow buffer	5 µl
1 mM dATP	10 µl
5U/ µl Klenow 3'- exo	3 µl

3. Incubate for 30 minutes at 37 °C in a hot block. (Lid must not exceed 50dC)

4. Clean up using SPRI beads (90µl beads, further as above) eluting in 16 µl of Sigma-water.

Ligation of Adapters to DNA Fragments

This protocol ligates adapters to the ends of the DNA fragments, in a molar ratio of 10:1, adapter to DNA insert, based on a starting quantity of 5 µg of DNA before fragmentation.

1. Mix in 1.5-ml tube:

PE-Adapter mix	10 µl
A-tailed DNA frags	15 µl
2x T4 DNA ligase Buffer	25 µl
H20	5 µl
DNA Ligase	5 µl

→ *Mix and Spin down*

2. Incubate for 15 minutes at 20°C.

3. clean up using SPRI beads (90µl beads, further as above) as described above, eluting in 51 µl Sigma-water.

pre-hyb PCR

Performing a small number of PCR cycles before hybridisation can improve robustness, particularly for clinical samples, and will simplify sample indexing. Amplify each 50 µl adapter-ligated library by dividing between 4 PCR reactions.

Herculase II for 4 cycles
Test with small aliquot for 4, 5 and 6 cycles before amplifying the large library!

For 10ul DNA

DNA template	10 µl
H2O	4.3 µl
PCR primer F	0.4 µl
PCR primer R	0.4 µl
5x Buffer	4 µl
dNTPs 10mM	0.5 µl
Herculase II	0.4 µl

PCR program

95°C	30s
95°C	10s
65°C	30s
72°C	30s
72°C	5 min
10°C	forever

SPRI bead cleanup of DNA library

Allow SPRI beads to come to room temperature for at least 30 minutes. Reagents need to be mixed well prior to use and should appear homogeneous and consistent in colour.

1. Add 90 µl of SPRI beads per 50 µl of adapter ligated sample in a 1.5 ml Lo-bind Eppendorf tube.

2. Vortex and leave at room temperature for 5 minutes.

3. Place tubes in a magnetic rack.

4. Leave for 5 minutes or until sample is clear.

5. Carefully remove the clear solution from the tubes and discard.

6. Dispense 700 µl of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.

7. Repeat the ethanol wash once again (total of two washes).

8. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 to 10 minutes or until the residual ethanol has evaporated.

9. Add 50 µl of molecular biology grade water, vortex and incubate at room temperature for 2 minutes.

10. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.

11. Carefully remove the water and retain in a new 1.5 ml lo-bind Eppendorf tube.

12. Repeat step 9-12 once more, retaining the water in the same 1.5 ml lo-bind tube. Total volume of elute should be 100 µl.

13. Centrifuge the eluate at 13,000 rpm in a bench top centrifuge for 10 minutes

14. Transfer the sample to a new 1.5 ml lo-bind tube leaving behind any precipitated beads.

15. Quantify 1 µl of the library using an Agilent DNA 1000 chip on a Bioanalyzer 2100 and proceed to hyb, following the manufacturer's recommended protocols.

Hybrid Capture protocol

Solution hybridisation (Agilent SureSelect protocol)

This protocol is for hybridization of adapter-ligated or PCR-amplified library DNA, so must be performed after Hybrid Capture Protocol 1.

Sample preparation

1. To block the reaction:

Prepped library	250ng
PCE Roche	5 µl
Hyb-Block #3	0.6 µl

Vortex → Spin down → dry in SpeedVec 45dC

Resuspend in H2O 4.5 µl

2. Put on Ice

3. Prepare Hybridization Buffer as follows. Volume for 1 capture:

SureSelect Hyb #1	25 µl
SureSelect Hyb #2	1 µl
SureSelect hyb #3	10 µl
SureSelect Hyb #4	13 µl

Note: *Do NOT keep on ice*.

4. Incubate 40 µl of Hybridization Buffer and library seperately at 95 °C for 5 minutes and 65 °C for at least 5 minutes. Keep at 65 °C until RNA baits are prepared (see below).

5. Dilute 0.5 µl RNAse Block with 0.5 µl Sigma-water

6. Add 0.5 µl diluted RNAse Block to 2.5 µl (250 ng) of RNA baits.

7. Incubate for 2 min at 65 °C.

Hybridization

Mix together in thermal cycler:

Hybridization Buffer	6.5 µl
RNA baits	3 µl
→ mix	
DNA library	4.5 µl
	14 µl

seal the tube and incubate **36 hours** at **65°C** with a heated lid at 105 °C in PCR machine

Prepare magnetic beads

1. Prewarm SureSelect Wash Buffer #2 at 65°C in Thermomixer

2. Vigorously resuspend the Dynal (Invitrogen) magnetic beads on a vortex mixer. Dynal beads settle during storage.

3. For each hybridization, add 50 µL Dynal magnetic beads to a 1.5 mL tube.

4. Wash the beads:

a, Add 200 µL SureSelect Binding buffer.
b, Mix the beads on a vortex mixer for 5 seconds.
c, Put tubes into a magnetic device, such as the Dynal magnetic separator (Invitrogen).
d, Remove and discard the supernatant.
e, Repeat step a through step d for a total of 3 washes.

5. Resuspend the beads in 200 µL of SureSelect Binding buffer.

Select hybrid capture with SureSelect

1. Preheat 20 µl Hybridisation Buffer to 65°C
2. Add 14 µl to hybridization mixture
3. Add the hybridization mixture directly from the thermocycler to the bead solution, and invert the tube to mix 3 to 5 times.
4. Incubate the hybrid-capture/bead solution on a Nutator for 30 minutes at RT

   → Make sure the sample is properly mixing in the tube.
5. Separate the beads and buffer on a Dynal magnetic separator and remove the supernatant.
6. Resuspend the beads in 500 µL SureSelect Wash Buffer #1 by vortexing 5 seconds
7. Incubate the samples for 15 minutes at RT, vortex in between
8. Wash the beads:

   a, Separate the beads and buffer on a Dynal magnetic separator and remove the supernatant.

   b, Mix the beads in prewarmed 500 µL SureSelect Wash Buffer #2 on a vortex mixer for 5 seconds to resuspend the beads.
   c, Incubate the samples for 10 minutes at 65°C. Vortex in between.

   d, Briefly spin in centrifuge.

   e, Repeat entire steps a-d for a total of 3 washes.

   → Make sure all of the wash buffer has been removed.
9. Mix the beads in 50 µL SureSelect Elution Buffer on a vortex mixer for 5 seconds to resuspend the beads.
10. Incubate the samples for 10 minutes at RT. Vortex in between.
11. Briefly spin in centrifuge.
12. Separate the beads and buffer on a Dynal magnetic separator.
13. Use a pipette to transfer the supernatant to a new 1.5 mL microcentrifuge tube.

➔ Contains the captured library!!! (discard the beads)

14. Add 50 µL of SureSelect Neutralization Buffer, briefly mix on vortex.

Desalt capture solution using SPRI beads

1. Let the SPRI XP beads come to room temperature for at least 30 minutes.

2. Mix the reagent well so that the reagent appears homogeneous and consistent in color. *Do not freeze*.

3. Add 180 µL of homogenous SPRI XP beads to a 1.5-mL LoBind tube, and add 100 µL of captured DNA library. Mix well on a vortex mixer and incubate for 5 minutes.

4. Put the tube in the magnetic stand. Wait for the solution to clear (approximately 3 to 5 minutes).

5. Keep the tube in the magnetic stand. Do not touch the beads while you carefully discard the cleared solution from the tubes.

6. Continue to keep the tube in the magnetic stand while you dispense 0.5 mL of 70% ethanol in each tube.

7. Let the tube sit for 1 minute to allow any disturbed beads to settle, and remove the ethanol.

8. Repeat step 6 and step7 step once.

9. Dry the samples on the 37°C heat block for 5 minutes or until the residual ethanol completely evaporates.

10. Add 30 µL nuclease-free water, mix well on a vortex mixer, and incubate for 2 minutes at room temperature.

11. Put the tube in the magnetic stand and leave for 2 to 3 minutes, until the solution is clear.

12. Remove the supernatant (~30 µL) to a fresh 1.5-mL LoBind tube. You can discard the beads at this time.

Hybrid Capture protocol 3
Post-hyb solution capture eluates

This protocol is for post-elution amplification of captured DNA

PCR primers:
PCR-F = 5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC 3′
PCR-R = 5′ CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC 3′

concentrate enriched library to ~15 µl;

1. Prepare the PCR master mix for Test PCR 10-12 cycles:

DNA template	3 µl
H2O	4.55 µl
PCR primer F	0.2 µl
PCR primer R	0.2 µl
5x Buffer	2 µl
dNTPs 10mM	0.25 µl
Herculase II	0.2 µl

PCR program

95°C	30s
95°C	10s
65°C	30s
72°C	30s
72°C	5 min
10°C	forever

4. Transfer PCR product into a 1.5ml Lo-Bind tube. Run the sample on an Agilent DNA 1000 chip on a Bioanalyzer 2100.

SPRI bead cleanup

Allow SPRI beads to come to room temperature for at least 30 minutes. Reagents need to be mixed well prior to use and should appear homogeneous and consistent in colour.

1. Take 90 µl of SPRI beads and add them to the 50 µl of PCR sample in a 1.5 ml Lo-Bind tube.

2. Vortex and hold at room temperature for 5 minutes.

3. Place tube in the magnetic rack and leave for 5 minutes or until sample is clear.

4. Carefully remove the clear solution from the tubes and discard.

5. Dispense 500 µl of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.

6. Repeat the ethanol wash once again. Total of two washes.

7. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 – 10 minutes or until the residual ethanol has evaporated.

8. Add 50 µl of molecular biology grade water, vortex and incubate at room temperature (20 °C) for 2 minutes.

9. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.

10. Carefully remove the water and retain in a new 1.5 ml Lo-Bind tube.

11. Repeat step 9 -12 once more, retaining the water in the same 1.5 ml Lo-Bind tube. Total volume of elute should be 100 µl.

12. Put the tube into the magnetic tool for 10 min.

13. Transfer the sample to a new 1.5 ml Lo-Bind tube leaving behind any precipitated beads.

After SPRI clean up, quantify by qPCR and sequence.

Annex 3 Malcolmson's 1-9 scale of increasing resistance scores. Score 9 as 0% necrotic tissue is not shown in the original figure from Cruickshank et al. (1982)

Thesis Florian Jupe

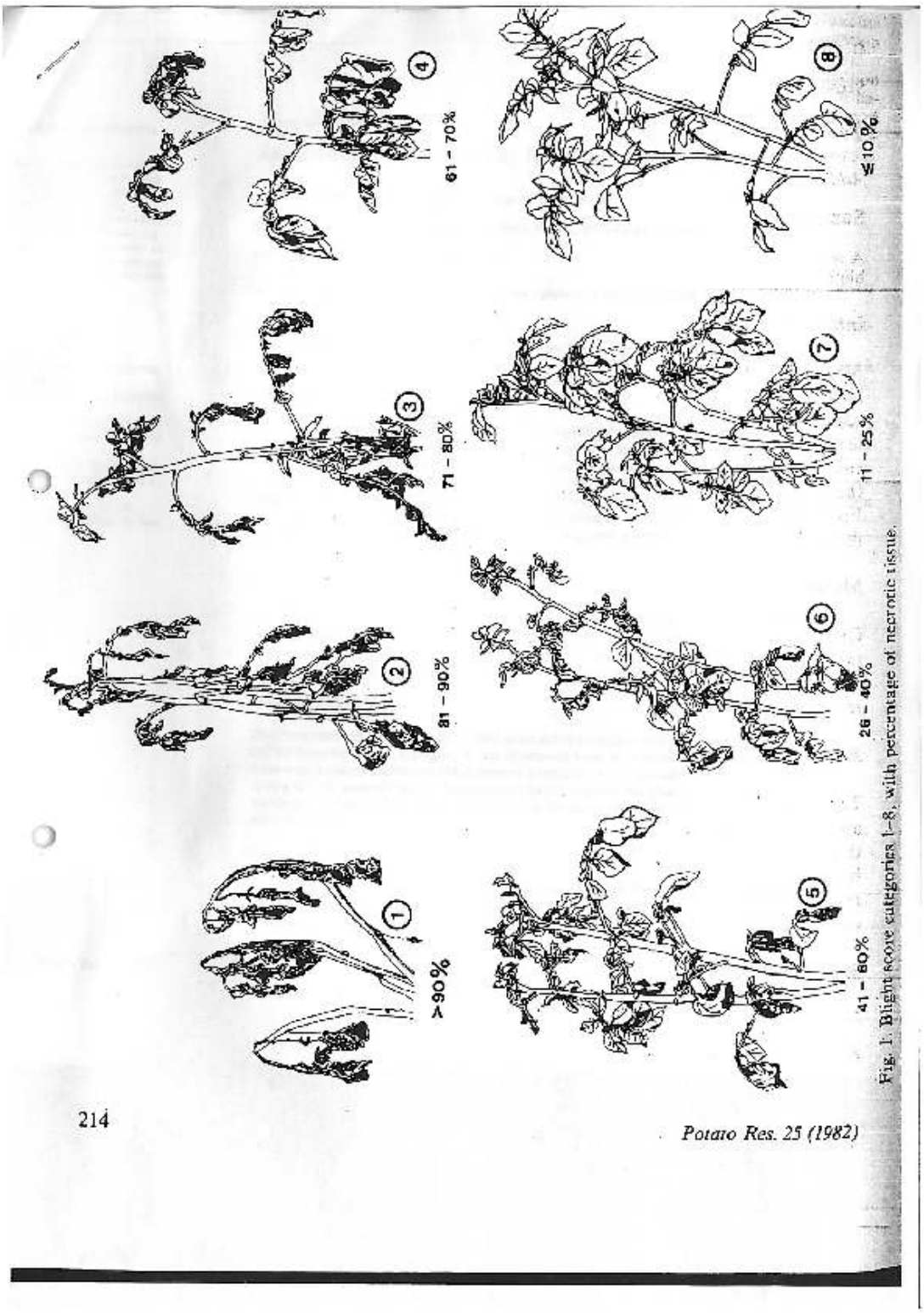


Fig. 1. Blight score categories 1–8, with percentage of necrotic tissue.