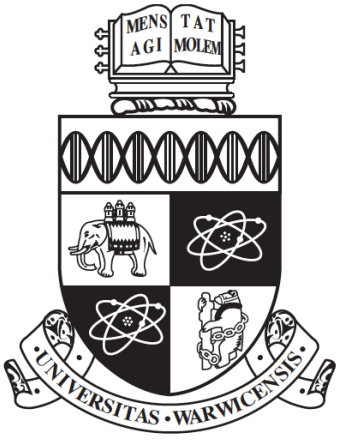
Advancing genomics to enable anticipatory breeding of resistance in *Brassica* crops to diseases of global importance

William Crowther



A thesis submitted to
The University of Warwick
For the degree of
Doctor of Philosophy in Life Sciences

School of Life Sciences
The University of Warwick
February 2021

List of Figures

List of Tables

Acknowledgements

I would first like to thank my supervisor Eric Holub, who bought me into his lab five years ago and developed me into a plant pathologist, a geneticist, a lover of beans and an all-round better scientist. His enthusiasm for the subject is truly infectious, with no resistance required. I would also like to thank my secondary supervisor Dr Charlotte Allender for her kind support as well as for provision of genebank seed stocks which were so pivotal to this project. Bioinformaticians Laura Baxter, Richard Stark, Andrew Armitage and George Crowther who provided me with fantastic support and guidance throughout my data analysis, each showing great deals of patience to help me understand and implement my coding. Particular thanks goes to Dr Volkan Cevik for his assistance in assembling pathogen genomes and guidance throughout their analysis. I'd also like to thank Joana Vincente, Sebastian Fairhead and Vania Horta de Passo for imparting their extensive knowledge of the subject, as well as for their advice and assistance with plant work. Thanks to everyone at the Warwick Crop Centre for keeping me relatively sane through the years, and in particular to Rosanne Maguire for her silliness, her crochet and plethora of bean related humour.

Finally, I'd like to say thank you to BBSRC and MIBTP for funding this research and providing me with such a fantastic opportunity to produce this research.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented was carried out by the author except in the cases outlined below:

- Sections 3.2.3 + 3.2.4, CRISPR construct assembly and the production of edited *Brassica oleracea* line DH1012 was undertaken by Tom Lawrenson at the Biotechnology Resources for Arable Crop Transformation (BRACT) group at the John Innes Centre in Norwich. This included genotyping of T_0 seedlings.
- The photographs in Figures 3.3, 3.4 and 3.5 were taken from the online BRACT JIC.ac.uk pictorial protocol (https://www.jic.ac.uk/app/uploads/2018/11/Brassica-Transformation-in-photos.pdf), with permission from the author.
- Section 4.3.2, the processing of Oxford Nanopore Technology raw reads and *de novo* assembly of AcBj12 and AcBjDC genomes was performed by Dr Volkan Cevik at the University of Bath, who also provided files detailing annotated CCG effectors.
- The photographs in Figures 1.2, 2.2, 2.5, 2.6 and 3.8 were taken by Professor Eric Holub.
- Genotyping by Sequencing in Section 2.2.5 was performed by LGC Genomics facility in Berlin, Germany.
- Whole genome resequencing in Section 2.2.6 was performed by Novogene (UK) in Cambridge.

Abstract

White rust (WR) (caused by the oomycete *Albugo candida*) inflicts significant yield decline in *brassica* crops across the world. It is a particular problem in countries like India, where resource poor farmers cultivating oilseed mustard (*Brassica juncea*) suffer between 30-60% yield losses per year. Breeding for host resistance is the most effective and environmentally friendly way of protecting *brassica* crops from WR. Rapid evolution of the pathogen highlights the need to efficiently map multiple resistance genes for combined integration into cultivars. Here we present novel methods for achieving this, with an emphasis on directly mapping from undeveloped genebank populations.

In this work, the application of Whole Genome Resequencing (WGS) and Bulked Segregant Analysis (BSA) proved to be highly effective for rapidly mapping sixteen major effect white rust resistance (WRR) QTLs across five geographically diverse *Brassica rapa* genebank populations, including a landrace and a wild species. This includes a QTL mapped on chromosome 6, containing a WRR allele that provides resistance to Donskaja-virulent UK race 2 isolate AcBjDC. WGS for all QTLs has been scrutinised, allowing identification of various polymorphic candidate WRR genes.

WRR was also explored in the *Brassica oleracea* EBH527 x A12DH mapping population, where a previously fine-mapped ACA2 locus contains a recessive race non-specific resistance. A CRISPR-Cas9 knockout of the primary candidate (a GDSL Lipase) yielded no alteration to susceptibility and WGS of the two parents has since been used to identify other potentially causal mutations within the region. WGS-BSA was also applied here to map a dominant resistance locus in tight-linkage to ACA2 from segregating recombinant line EH177.

Finally, Oxford Nanopore Technology long-read sequencing was used to produce high-quality genome assemblies for race 2 isolates AcBj12 and AcBjDC, augmented with Illumina reads. The AcBj12_ONT assembly provided a 30 % improvement in contiguity relative to previous efforts using PacBio data. Comparative genomics that included Donskaja-virulent isolate Ac2v was used to document allelic variation in “CCG” class effectors that provides a platform for future identification of the *Avr* elicitor. Collectively, methods presented here enhance rapid identification of loci and candidate genes for the benefit of global brassica production.

List of Abbreviations

Ac	*Albugo candida*
AcBoWells	*Albugo candida* isolate from Wellesbourne, UK
AcBj12	*Albugo candida* isolate from UK mustard production
AcBjDC	*Albugo candida* isolate from UK mustard production
Ac2v	*Albugo candida* isolate from Canada
AF	Allele frequency
Avr	Avirulence
BSA	Bulked Segregant Analysis
BWA	Burrows-Wheeler Aligner
CAPS	Cleaved amplified polymorphic sequence
CC	Coiled-coil
CCG	CxxCxxxxxG amino-acid motif
CNL	CC-NB-LRR
Col	*Arabidopsis thaliana* accession Columbia
CDS	Coding sequences
DH	Doubled haploid
DNA	Deoxyribonucleic acid
DPI	Days-post-inoculation
ECM	Extra cellular matrix
EDS	Enhanced disease susceptibility
ERF	Ethylene responsive transcription factor
ET	Ethylene responses
ETI	Effector-triggered immunity
FDR	Benjamini-Hochberg false discovery rate
GBS	Genotyping by Sequencing
HR	Hypersensitive response
HWE	Hardy-Weinberg equilibrium
HMM	Hidden Markov modules
ID	Integrated domain
INDEL	Insertion / deletion polymorphism

JA	Jasmonic acid
LD	Linkage disequilibrium
LM	Linkage mapping
LRR	Leucine rich repeat
MAGIC	Multiparent Advanced Generation Inter-Cross
MAS	Marker assisted selection
MAF	Minimum allele frequency
NB	Nucleotide binding
NDR	Non-race-specific disease resistance
NGS	Next generation sequencing
NLR	Nucleotide-binding, leucine-rich-repeat
PAMP	Pathogen associated molecular pattern
PCR	Polymerase chain reaction
PenSeq	Sequence capture-based protocol
PGR	Plant genetic resources
PRR	Pattern-recognition receptor
PTI	PAMP-triggered immunity
QTL	Quantitative trait loci
R-gene	Resistance gene
raSNP	Resistance-associated SNPs
RLK	Receptor like kinase
RLP	Receptor like protein
RenSeq	Resistant gene enrichment and sequencing
RIL	Recombinant inbred lines
SA	Salicylic acid
SNP	Single nucleotide polymorphism
STPK	Serine/threonine-protein kinase
TAL	Transcription activator-like
TD	Exon tandem duplication
TIR	Toll/interleukin receptor
TNL	TIR-NB-LRR
WGS	Whole genome resequencing

WGD	Whole genome duplication
WGT	Whole genome triplication
WRR	White rust resistance

Chapter 1

General Introduction

1.1 Preface

"You can't build a peaceful world on empty stomachs and human misery" – *Norman Borlaug*

The remainder of the century and beyond will present the greatest challenges faced in the history of crop development and global food security. Human populations are projected to reach 9.8 billion people by 2050, which will require an increase in agricultural output of 60%, as well as increased provision of balanced nutritional diets within developing countries (FAO, 2016). This target is made even more challenging by an approaching 'perfect storm' of global issues, including the decreasing availability of farmland per capita, water and energy scarcity, as well as the increasing impacts of extreme weather events associated with climate change. These environmental perturbations are impacting the global distributions and severity of pests and pathogens, whose traditional control by pesticide application is becoming increasingly restricted by tightening agrochemical legislations. This leaves farmers with a limited ability to protect their crops from disease.

Put simply, agriculture must rapidly increase outputs, without further conversion of non-agricultural land and whilst simultaneously reducing inputs of water, energy and chemicals. This requires sustainable intensification, where increases in yields do not come at a cost to biodiversity and ecosystem services upon which food systems are ultimately dependent. Failure to do so will result in widespread food insecurity, ecological destruction, civil conflict and human suffering.

The goals of sustainable intensification necessitate an unprecedented transformation in global food systems akin to the next agricultural revolution. This will require vast global efforts of innovation across both biotic and abiotic components of agriculture; from enhancing the natural ecological processes underpinning food systems, to utilising agri-tech for production and distribution of harvests with reduced wastage. Ultimately, this will depend upon the rapid adoption of research-based strategies that enhance both the productivity and resilience of agroecosystems in the light of impending environmental stresses.

1.2 Founding fathers for modern use of plant genetic resources

Plant genetic resources (PGRs) are an invaluable foundation for the genetic improvement of crops, providing allelic diversity for breeding of desirable traits such as high yields, environmental adaptation, and nutritional quality for human consumption. However, continual selection for such traits over several millennia has resulted in genome-wide divergence in crops relative to their wild progenitors, with a broad trend towards a reduction of genetic diversity within the crop species. This narrow genetic basis of most crop species today has direct implications for their ability to withstand environmental perturbations such as climate extremes as well as their ability to resist damage from pathogens and insect pests. Genetic improvement of crop varieties is a global agricultural priority for meeting sustainability targets, with an emphasis on utilizing the natural allelic diversity contained within plant genebanks. Integration of disease resistance traits from diverse stores of germplasm into crop varieties can best be understood within a historical context of the subject.

The Austrian monk Gregor Mendel (1822 - 1884) discovered the theoretical foundations of genetics through his seminal work on the common pea (*Pisum* sativum) to "deduce the law according to which they (traits) appear in successive generations" (Mendel, 1901). Using pure-breeding varieties of pea as an experimental system, Mendel produced monohybrid crosses to observe inheritance for seven traits (phenotypes) including: flower colour, mature seed colour and form (wrinkled or smooth), mature pod colour and form (inflated or constricted), colour of the unripe pod and overall plant height. His key findings were that F_1 hybrid (first generation) offspring all resembled one of the two parental phenotypes, with no 'blending' of the trait (as was generally expected at the time). Also, amongst the offspring derived from F_1 hybrids, both the parental phenotypes were observed, suggesting that hybrids retain two factors (now referred to as alleles) for producing both parental types. He described the trait observed in the F_1 hybrid as dominant, and the alternative trait that reappears in offspring of the following generation as recessive. By conducting hundreds of crosses and extensively counting the numbers of individual offspring from each phenotype class, Mendel was able to observe an approximate ratio of 3:1 for dominant vs recessive in the second-generation offspring. From his data, Mendel surmised his law of segregation; that the two phenotypes observed consisted of three classes; a dominant (AA), a recessive (aa) and a hybrid (Aa), where the

hybrid always displays the dominant phenotype. When exploring results from dihybrid crosses (where parents differ in two observable traits) Mendel found that the resulting segregation ratio of 9:3:3:1 fit the product of two 1:2:1 ratios. He concluded that both traits segregate independently, where inheritance of one characteristic does not affect inheritance of the other, a law he termed independent assortment. Whilst the significance of Mendel's findings were not realised for more than three decades after publishing in 1865, they ultimately provided a paradigm shift in understanding of heredity and genetics.

The Russian plant geneticist Nickolai Vavilov (1887 - 1943) was greatly influenced throughout his life's research by established Mendelian and Darwinian theories. Vavilov was a plant breeder and head of the All-Union V.I. Lenin Academy of Agricultural Sciences, where he founded over 400 research institutes throughout Russia (Cohen & Loskutov, 2016). He was also an avid collector of plants and during numerous overseas expeditions Vavilov amassed a collection of approximately 50,000 wild varieties. Observations made during this time lead Vavilov to postulate that a crop species evolutionary centre of origin would be from regions where the wild progenitor species showed maximum adaptiveness and diversity.

Before his time, Vavilov saw the potential of non-cultivated diversity as a means to improve current crops by providing plant breeders with the necessary genetic variation from which to produce new varieties with qualities such as disease resistance and tolerance to adverse growing conditions. From his travels he also became aware of the growing threat to this wild diversity posed by human activities and the modernisation of agriculture, a phenomenon he described as "genetic erosion" (Hummer, 2015). Subsequently, Vavilov formed the first *ex situ* genebank in the world in St Petersberg in the 1920s (today called the N.I. Vavilov Research Institute for plant industry) with the aim of preserving wild species diversity and accessions. Despite the theoretical and practical contributions Vavilov made to crop science, species diversity and conservation, political leaders in Russia came to reject Mendelian principals and ultimately Vavilovs research. Tragically, this resulted his incarceration in 1941, where he two years later he died of likely starvation. Vavilov was eventually pardoned posthumously in 1955 by the Russian government and today is fully recognised for his contributions to science. His work and philosophy provided a great inspiration for the theoretical direction of this thesis.

The idea that Mendelian principals could be used to harness key agricultural traits such as disease resistance from natural plant genetic resource collections was first demonstrated by Rowland Biffen in 1907, through his breeding of yellow rust (*Puccinia striiformis f.sp. tritici*) resistance in wheat (Biffen, 1907). Biffen observed that crosses between the partially resistant cultivar 'Rivet', and the highly susceptible cultivar 'Red King' produced F_1 hybrids that all showed Red King susceptibility. The F_2 generation produced 194 susceptible to 64 rust-free (resistant) individuals in an approximate 3:1 Mendelian ratio. Data from both generations indicated that resistance was recessive, or that susceptibility is conferred as a dominant trait. From his assessment of F_3 lines, Biffen found that offspring from partially resistant individuals were fixed for the trait (F_2 was homozygous for a resistance allele), whilst those from susceptible F_2 parents either conferred full susceptibility (F_2 was homozygous for a susceptibility allele) or segregated in a 3:1 ratio (F_2 was heterozygous). Importantly, using his understanding of wheat yellow rust resistance as a single recessive gene, Biffen was then able to deploy of the resistance trait into new wheat varieties, which went on to provide longstanding (durable) control of yellow rust.

In the mid 1900's, the plant pathologist/geneticist Harold Flor went on to extend the work of Biffen and Mendel by applying genetics to the investigation of corresponding traits in both the host plant (resistant and susceptible) and the pathogen (avirulent and virulent). He used the experimental pathosystem of flax or linseed (*Linum usitatissimum*) and the fungal rust (*Melampsora lini*) (Flor, 1942; Flor, 1947; Flor 1955). Flor was adept at producing rust hybrids through cross-fertilisation of different pathotypes which he used to screen F_2 progeny of differential flax genotypes that were each homozygous for a different resistance allele. In the pathogen, Flor observed 3:1 ratios of avirulent to virulent progeny and was able to conclude that a single avirulence (*avr*) allele in the pathogen corresponded to a matching resistance (*R*) allele, whose interaction resulted in an incompatible phenotype (no disease). The term 'gene-for-gene' was coined to characterise the inheritance of matching resistance and avirulence pairing to explain rust resistance in flax (Flor, 1971). Flor consciously excluded examples of alternative inheritance for host-pathogen interactions from his pathosystem (*e.g.,* digenic inheritance and recessive resistance) to simplify his findings, even though he was certainly familiar with seminal work of Biffen on recessive yellow rust resistance in wheat. However, a

simplified theory was needed to present plant pathologists with tractable examples for a genetical approach to research of disease resistance.

'Gene-for-gene' was borne and provided a testable hypothesis for investigating the molecular basis of plant-pathogen interactions. Albert Ellingboe was an early geneticist who hypothesised that a direct interaction was occurring between a product of the pathogen race-specificity allele and the complementary *R*-allele product in the host (Ellingboe, 1976). This provided seminal biochemical predictions of pathogen receptor-like proteins as the basis for *R*-gene mediated triggering of plant defence, now referred to as the plant innate immune system.

1.3 Molecular basis of induced host defence and the innate immune system of plants

Pathogens capable of breaching the external waxy cuticle or "skin" layer of the plant encounter an active immune system which is highly adapted to detect pathogen molecules produced during infection (Amir *et al*., 2008; Chisholm *et al*., 2006; Dodds & Rathjen, 2010; Jones, 2006). Host receptor-like proteins embedded in the plasma membrane enable detection of pathogen molecules released in the apoplast, and cytoplasmic receptors enable detection of pathogen molecules released within the host cell. Detection then triggers a rapid cascade of kinase-mediated defence responses including programmed cell death (often referred to as a hypersensitive response or HR) of the penetrated host cell (Balint-Kurti, 2019), and hormonal induction of broad-spectrum systemic acquired resistance in surrounding cells (Ryals, 1996; Dangl & Jones, 2001). This effectively quarantines biotrophic pathogens from acquiring their necessary substrate, and thereby arresting further proliferation in host tissue. It is now know that the *R*-genes predicted by Flor and Ellingboe are explained by allelic variation in genes encoding different classes of receptor-like proteins. These genes are explained in more detail below. Similarly, predicted avirulence (Avr) proteins that elicit defence when detected by a corresponding *R*-protein have been characterised from a wide range of pathogens.

Interestingly, AVR proteins generally provide a role as effectors of pathogenicity during infection of a susceptible host genotype. All major classes of plant pathogens including viruses, fungi, oomycetes, bacteria, nematodes, and feeding insects have evolved vast repertoires of highly diverse effector proteins (Baltrus *et al*., 2011; Quispe-

Tintaya, 2017; Raffaele *et al*., 2010). Pathogen effectors are under continuous selection pressure to facilitate compatible (virulent allele in a susceptible host) rather than incompatible (avirulent in a resistant host) interactions. Compatibility can result from effectors that either subvert defence responses or facilitate acquisition of nutrients.

The development of recombinant DNA technology was pivotal for molecular plant pathology, allowing further understanding of host-pathogen mechanistic interactions. For example, the first *Avr* elicitor was cloned from a race 6 isolate of *Pseudomonas syringae pv. glycinea* which is the causal agent of bacterial blight of soybean (Staskawicz *et al*., 1984). A library of 680 randomly cloned genomic fragments were individually conjugated to a race 5 isolate of *P. s. glycinea* and inoculated onto susceptible soybean cultivars. From these, a single clone (pPg6L3) was responsible for changing the wild-type virulent phenotype of the race 5 isolate from virulent to avirulent, producing a HR response in the host plant. The result stimulated wider research efforts to clone and identify *Avr* elicitors in other groups of pathogen including *Avr9* in the fungal species *Cladosporium fulvum* (Van Den Ackerveken *et al*., 1993) and *Avr1b* in the oomycete *Phytophthora sojae* (Shan *et al*., 2004).

The first *R*-gene was molecularly characterised in 1993 as *Pto* from tomato which encodes a cytoplasmic kinase (Martin *et al.* 1993). The *Pto* gene cluster was identified from a yeast artificial chromosome clone, where a cDNA clone of *Pto* provided resistance to any race of *P. syringae* pv. *tomato* that carries the previously identified *Avr* effector *AvrPto* (Ronald *et al*., 1992). Later, a study using yeast two-hybridisation experiments confirmed the physical interaction of the *Pto* and *AvrPto* proteins, providing the molecular evidence of both components of the gene-for-gene hypothesis (Tang *et al*., 1996).

Since the discovery of *Pto,* several classes of *R*-genes encoding receptor-like proteins have been molecularly characterised in a range of plant species. Extracellular pattern-recognition receptors (PRRs) include receptor like kinases (RLK) and receptor like proteins (RLP) that are embedded in the plasma membrane and confer function by detecting pathogen-associated molecular patterns (PAMPs) such as bacterial flagellin or fungal chitin (Zipfel, 2008). The high conservation of PAMPs across whole classes of microbes suggests they are indispensable components of the pathogen, and therefore difficult to mutate. PAMP detection by PRRs is referred to as PAMP-triggered immunity (PTI) which triggers intracellular signalling and transcriptional regulation of defence

responses such as release of reactive-oxygen species and callose deposition for cell wall reinforcement (Bailey-Serres & Mittler, 2006; Luna *et al*., 2011)

Within plant cells, cytoplasmic *R*-proteins detect pathogen effectors either directly or indirectly via effector activity on the host substrate (Chisholm *et al*., 2006). Detection instigates effector-triggered immunity (ETI) and leads to rapid induction of HR in locally infected tissues (Jones, 2006). *R*-genes represent a class of cytoplasmic receptors termed NLRs, that typically include a central nucleotide binding (NB) domain, a C-terminal leucine-rich repeat (LRR) and a N-terminal signalling domain (Meyers *et al*., 2003). NLRs are grouped into two subclasses, containing either a coiled coil (CC) or a human toll interleukin receptor (TIR) signalling domain (Meyers *et al*., 2002, 2003) where the activity of each class is dependent on different network components. The class TIR-NB-LRR (TNL) receptors require a functional copy of an enhanced disease susceptibility gene (*EDS1*) whilst the CC-NB-LRR (CNL) class often requires the non-race-specific disease resistance (*NDR1*) to induce a defence response, representing two distinct resistance pathways (Aarts *et al.*, 1998).

The extensive diversity of NLR genes, both within and between species is becoming increasingly well documented (Wei *et al*., 2016), where copy number variation has been shown to differ by orders of magnitude. For example, apple (Malus x domestica) has a reported 1015 NLR genes (Arya *et al*., 2014), whilst papaya (*Carica papaya*) only contains 54 (Porter *et al*., 2009) which is a significantly lower number than species with smaller genomes such as *Arabidopsis thaliana* (159 NLRs) (Guo *et al*., 2011). Such copy number variation is formed within plant genomes through a combination of hybridisation, whole genome duplication (WGD), individual gene duplications, transposon activity and mutation. This produces functionally redundant gene copies that are free for selection of new NLR specificities, facilitating coevolution with pathogen effector repertoires. Specifically, exon tandem duplication (TD) is considered to be the primary mechanism for increasing gene copy number and is responsible for production of complex NLR clusters with variable domain architectures (Cannon *et al*., 2004; Yang *et al*., 2008). Examples have been found that lack LRR domains such as *RLM3* (TIR-NB) in *Arabidopsis* (Staal *et al*., 2008), or contain additional integrated domains (IDs) (Baggs *et al*., 2017). Discovery of NLR-IDs provided a major shift in understanding of NLR function as some exogenous domain fusions have evolved to resemble pathogen targets. The WRKY DNA-binding domain of

RRS1 is an example of this where interaction with the wilt pathogen *Ralstonia solanacearum* effector *PopP2* or the leaf pathogen *Pseudomonas syringae pv. Pisi* effector *AvrRps4* modifies the ID to trigger an immune response (Staal *et al*., 2008).

1.4 The *Brassicaceae*

The large angiosperm and dicot family the *Brassicaceae* (order Brassicales) constitutes nearly 338 genera and 3,709 species, distributed globally across every continent except Antarctica (Anjum *et al*., 2012). Several species have great agricultural, economic and scientific importance such as the diverse genus of *Brassica* crop species and the wild relative *Arabidopsis thaliana*.

The German PhD student Friedrich Laibach originally chose *A. thaliana* as a study species in 1907 (University of Bonn, Germany) and found the plant presented a number of useful characteristics for study. These include that it contains only a small number of chromosomes ($n = 5$), displays significant natural variation for a number of traits, produces fertile hybrids and rapidly produces large amounts of progeny that can be easily cultivated to adult plants in ~ 8 weeks (Jare & Williams, 2014). By the late 1980s *A. thaliana* had been widely adopted as the model species to study plant molecular biology. It became the first plant species put forwards for genome sequencing in the mid-1990s, where collaborative efforts of the *Arabidopsis* Genome Initiative culminated in the landmark publication of complete sequence and analysis in 2000 (Poczai *et al*., 2014). This resource has been vital for driving contemporary understanding of plant biology, which in turn has provided new means for developing crops that contribute to issues of food security.

The *Brassica* genus is comprised of 76 species which are all highly diverse and show equally variable morphotypes in terms of plant size, colour, leaf shape, hypocotyl, flower buds, branching patterns and seed oil composition (Nanjundan *et al*., 2020). This diversity has enabled selection and breeding of *Brassica* crops that are widely adapted to climates and environmental conditions around the world, with harvestable produce that includes oilseed, condiments, fodder and vegetable crops (Cheng *et al*., 2014). There are six major *Brassica* crop species including three diploid species (*B. rapa* = AA, $n = 10$; *B. nigra* = BB, $n = 8$; and *B. oleracea* = CC, $n = 9$) and three amphidiploids whose genomes are constituted by the diploid species (*B. juncea* = AABB, $n = 18$; *B. napus* = AACC, $n = 19$; and *B. carinata* = BBCC, $n = 17$). The relationships between these species were first

established in 1935 by Nagaharu U as the 'Triangle of U' (Figure 1.1.) who made extensive crosses between diploid and tetraploid plants and examined how the chromosomes paired in the resulting hybrids (U, 1935).

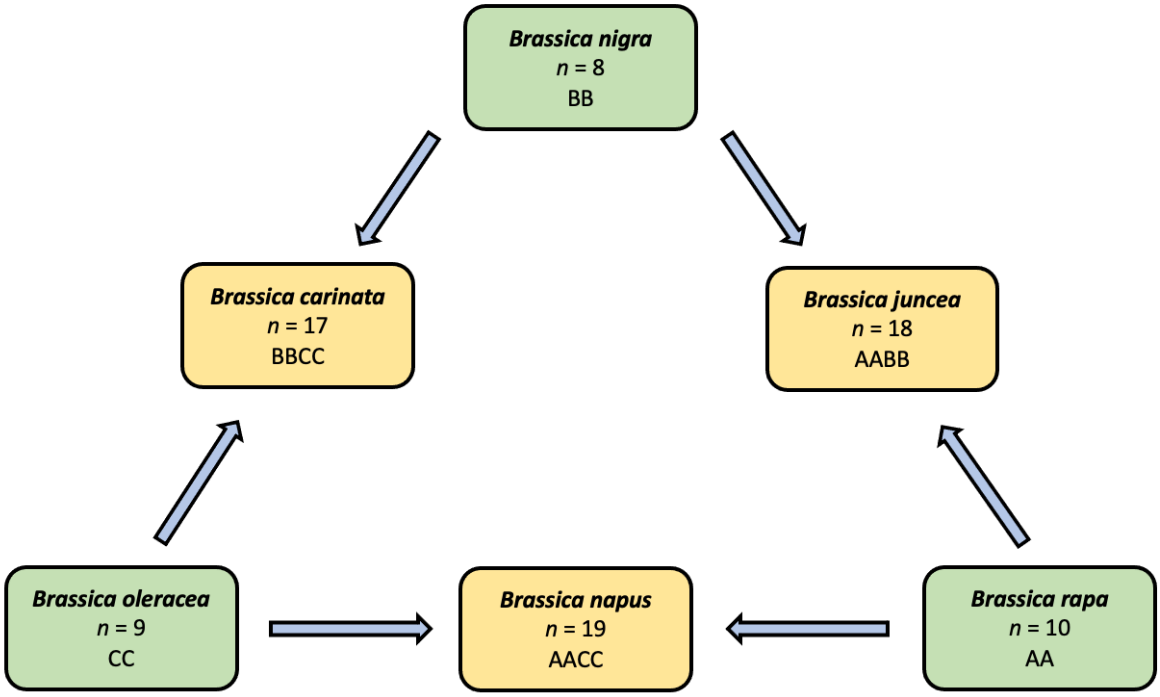


Figure 1.1. The triangle of U (U, 1935) describing the relationship between the three diploid *Brassica* crop species (*B. rapa*, A genome; *B. nigra*, B genome; and *B. oleracea*, C genome), and the intermediate hybrid or allotetraploid species (*B. juncea*, *B. napus* and *B. carinata*).

The *Brassica* lineage underwent a whole genome triplication (WGT) event after splitting from the *Arabidopsis* genera at a still disputed period approximately either 7.9–18.8 Mya years ago (Lysak *et al*., 2005) or 22.5 Mya (Beilstein *et al*., 2010). This event was important to the speciation and expansion of brassica crops, where the subsequent genomic rearrangement and gene evolution initiated by WGT resulted in the rich diversity of morphotypes found today. In the *Brassica* crop species this variation includes leafy heads, enlarged roots, variable stem or inflorescence structure, oilseeds, or even ornamental features. Many of these features have developed independently, such as the distinct leafy head of *B. rapa* which is also produced by *B. oleracea* and *B. juncea*. Whilst the enlarged root of *B. rapa* turnip varieties are also a feature of *B. juncea* and *B. napus*

(Anjum *et al*., 2012). *B. rapa* forms include turnip, napa cabbage, bomdong, bok choi and rapini. *B. nigra* is primarily cultivated for its dark coloured seed which is commonly used as spice. *B. oleracea*, commonly referred to as 'the dog of the plant world' due to its hyperdiversity, includes broccoli, Brussels sprout, cabbage, cauliflower, kale and kohlrabi varieties.

Brassica juncea is a mustard crop, with English, Indian and Chinese varieties. *B. caratinia* is less widely cultivated globally, though is grown in Ethiopia as an oilseed crop and for biofuel. *B. napus*, otherwise known as rapeseed, is cultivated for its oil-rich seed that contains low levels of erucic acid. Also bred for turnip varieties, *B. napus* is considered to be the third largest source of vegetable oil in the world and the second largest source of protein meal (Anjum *et al*., 2012). There are also numerous wild relative or lesser-known *Brassica* crop species that have highly diversified morphotypes. These represent a virtually untapped genetic resource for development of agronomic traits such as disease resistance (Warwick & Al-Shehbaz, 2006).

Brassica species diversity is mirrored by an equally diverse array of pathogen species that have coevolved to infect the various host types. Exploitation of host resources by pathogens and subsequent disease causes significant agricultural and economic losses, symptoms of which are frequently treated by application of environmentally damaging chemicals. Breeding for host resistance is considered to be the most efficient, cost-effective and environmentally sustainable way of protecting *Brassica* plants from disease. A number of *R*-genes have now been identified and introgressed into brassica crop species by use of conventional breeding efforts aided by marker assisted selection (MAS) or transgenic methods. This not only provides crop protection but has also significantly advanced understanding of host-pathogen interactions. Research efforts have been hugely enhanced since the 2010s with the widespread development and adoption of genomic and 'omics' technologies. The availability of these vast genetic datasets, and in particular the release of all six brassica crop species reference genomes, has been pivotal in the elucidation of *R*-genes and their functions (Chalhoub *et al*., 2014; Inturrisi *et al*., 2020; S. Liu *et al*., 2014; Parkin *et al*., 2014; J. Yang *et al*., 2016).

Turnip mosaic virus (TuMV) is the most prevalent virus that infects brassica crops, causing heavy economic crop losses of up to 30% across Europe, Asia and North America

(Walsh, 2002). High variation exhibited by TuMV, as well as its widespread nonpersistent dispersal by 89 species of aphid make the disease difficult to control. Fortunately, more than ten TuMV *R*-genes have now been characterised in brassica crops so far, predominantly from the A genomes of *B. rapa* and *B. napus* (Lv *et al*., 2020). Work is being undertaken to utilise TuMV *R*-genes in breeding programs, such as by use of Kompetitive Allele-Specific PCR (KASP) markers located within the TuMV recessive *R*-gene *retro2* for high-throughput MAS application (Li *et al*., 2016). Direct breeding methods have also been undertaken by transforming the susceptible *B. rapa* cultivar 'Seoul' with eukaryotic initiation factor eIF(iso)4E variants to produce broad resistance to multiple TuMV strains (Kim *et al*., 2014).

Black rot (BR), caused by *Xanthomonas campestris pv. campestris* (*Xcc*) is the major bacterial pathogen of brassica crops. Most BR research to date has focused on preliminary mapping of QTLs in the A genome to *Xcc* races 1 and 4. Though markers tightly linked to the Xca1Bo resistance locus in *B. oleracea* show potential for MAS cauliflower breeding (Kalia *et al*., 2017).

Black leg or stem canker is caused by the air borne fungal agent *Leptosphaeria maculans* (*Lm*), which is highly destructive to brassica crops in Australia, North America and Europe (Fitt *et al*., 2006). Extensive mapping of *Lm* resistance loci has been conducted since the 1990s and improved resistance cultivars are frequently produced today using MAS (Lv *et al*., 2020).

Stem rot, caused by the soil borne fungus *Sclerotinia sclerotiorum* (*Ss*) is devastating for brassica production and in particular for oilseed rape production where yield losses can be as high as 80%. Spores can persist in soil for several years to reinfect crops and no highly resistant sources of *Ss* resistance have yet been found, impeding on breeding efforts. Some moderate resistances have been mapped in *B. napus*, though research is now focusing on wild relatives where *Ss* resistance from *Brassica incana* has been bred unto *B. napus* using a combination of hexaploidy hybridization and MAS (Mei *et al*., 2015).

Fusarium wilt, caused by the fungal pathogen *Fusarium oxysporum f. sp. conglutinans* (*Foc*) provides a global threat to *Brassica* production. *B. oleracea* provides the main source for *Foc* resistance and gene targets such as FOC1 are currently being developed in breeding programs using MAS (Lv *et al*., 2014).

Downy Mildew is a foliar disease, caused by the oomycete *Hyaloperonospora brassicae* (*Hb*) that inflicts considerable yield losses to brassica crops particularly in Europe, Asia and Australia. Markers closely linked to several resistance loci in in *B. oleracea* and *B. rapa* are being utilised with MAS for improved resistance breeding. For example, Yu *et al.*, (2011) developed a sequence-characterized amplified region (SCAR) marker from the random amplified polymorphic (RAPD) marker K14-1030 to enable early selection in resistant progenies.

Clubroot is caused by the soil borne *Plasmodiophora brassicae* (*Pb*) from the taxon Rhizaria (Yu *et al*., 2011), and is rapidly becoming a major threat to *Brassica* production around the world. *Pb* is a highly variable pathogen whose ability to survive in the soil for many years as resting spores make it difficult to control. Efforts to identify resistance to *Pb* have been extensive and have collectively identified more resistance loci than any other brassica disease (Lv *et al*., 2020), with MAS applied to produce a series of resistant cultivars. For example, distant hybridisation embryo rescue and MAS have been used to transfer *B. rapa Pb* resistance loci into *B. napus* to produce clubroot resistant hybrids (Liu et al., 2018).

1.5 White Blister Rust

The eukaryotic oomycete *Albugo candida* belongs to the order Albuginales, which is comprised exclusively of obligate plant pathogens (Choi *et al*., 2008). This clade is part of the peronosporalean lineage that includes other major plant pathogens such as Phytophthora, Pythium and downy mildews (Thines & Kamoun, 2010). The genus *Albugo* contains more than 40 species that infect over 400 host species that span 31 families of dicots and one family of monocot (Biga, 1955; Choi *et al*., 2009; Erma, n.d.; Meena *et al*., 2014). Several species cause extensive yield losses to agriculturally important crops including *Albugo candida* on oil seed and vegetable brassicas, *A. tragopogonis* on sunflower, *A. ipomoeae-pandurate* on sweet potato and *A. occidentalis* on spinach (Saharan *et al*., 2014).

Albugo candida is the causal agent of the globally distributed white blister rust and can infect as many 63 genera and 241 species, including those in the Brassicaceae, Cleomaceae and Capparaceae families (Biga, 1955; Choi *et al*., 2009; Meena *et al*., 2014). Infections of economically important oilseed and vegetable crops worldwide can result in

losses of between 9 – 60 % (Harper, 1974; Ram & Awasthi, 2018; Saharan & Verma, 1992). The initial stage of infection is characterised by development of white pustules (1 – 2 mm in diameter) on foliage, which converge to form larger characteristic blisters (Figure 1.2). This impedes the photosynthetic capacity of the host and thereby reduces the nutritional and aesthetic quality of the harvested product. Systemic spread of hyphae through the plant can then produce stem blisters which reduces structural integrity of the plant as well as impede transpiration and transportation of nutrients. The most severe crop damage often results from pustules produced on the inflorescence, where the effects of hypertrophy and hyperplasia distort tissue growth to produce characteristic 'staghead' formations. This leads to abortion of siliques and substantial loss of harvestable seed. This is a major problem for oilseed growing regions of the world such as India where cultivars of *B. juncea* are highly susceptible to white rust, and where climatic conditions in the northern oilseed growing regions favour growth of the pathogen (Li *et al*., 2009). Favourable environmental conditions for *A. candida* proliferation include 12 - 15 °C temperature, > 70 % relative humidity and 2.7 – 3.4 km/h of wind velocity with intermittent rain (Saharan & Verma 1992; Chattopadhyay *et al*., 2011).

The white blisters that characterise infection by the diploid agent *A. candida* result from the exerted pressure generated within tissues by rapidly multiplying sporangiospore mother cells and linked chains of desiccated zoosporangia (Holub *et al*., 1995). This results in eventual rupturing of the sorus, where expelled zoosporangia are dispersed by air currents or rain droplets. Water is crucial for successful germination of zoosporangia, where rehydration stimulates cellular differentiation to produce four to six biflagellate zoospores. These become motile when released and navigate towards host stomatal pores where they encyst, losing their flagella and producing a cell wall. A short germ tube is then formed which extends into the substomatal chamber and penetrates adjacent mesophyll cells via a haustorial structure.

Where a host pathogen interaction is compatible then hyphae are produced that spread through the mesophyll to access host cells with production of further haustoria. Hyphae can then access plant vasculature to facilitate access to the hypercotyl and promote systemic spread throughout the plant. Production of new pustules and dispersal of zoosporangia represents completion of the asexual lifecycle of *A. candida*. Alternatively, sexual reproduction can occur through production of male and female

gametangia called antheridia and oogonia which develop at hyphal tips within the inflorescence tissue. Fertilisation leads to production of large oospores with thick cell walls that provide protection through a period of dormancy within the soil before germinating to release zoospores (Holub *et al.*, 1995). This sexual phase provides *A. candida* with both the means to persist through unfavourable environmental conditions as well as to genetically recombine, facilitating adaptation to numerous hosts.

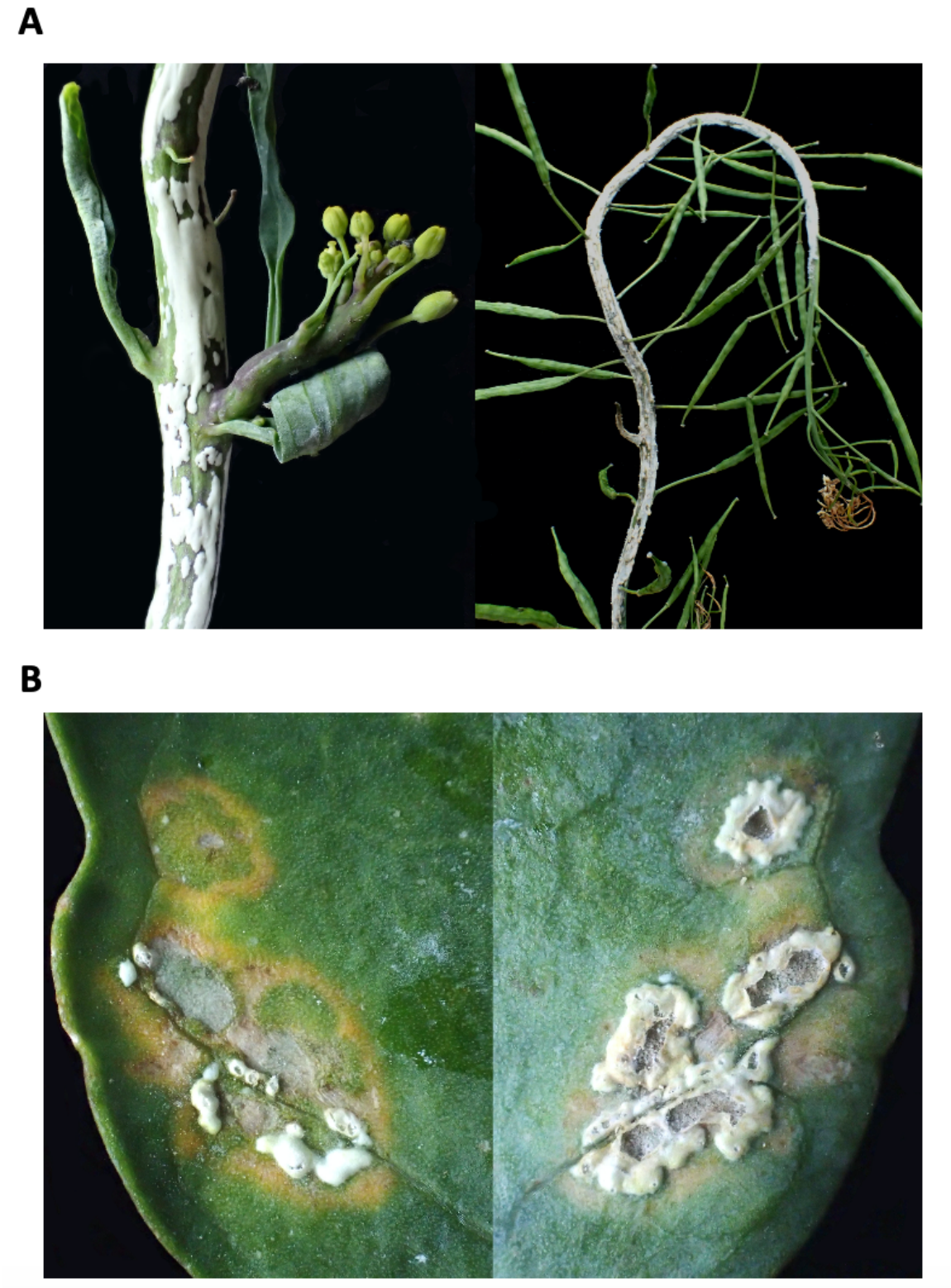


Figure 1.2. White blister rust caused by the biotrophic oomycete *Albugo candida* growing on: A); Stem and inflorescence of *B. juncea* having *infiltrated* systemically throughout the plant. B); leaf tissue of *Brassica juncea*

A. candida is a highly specialised pathogen that is currently classified into 17 distinct physiological races based on specificity to different species of *Brassica*. Pound and Williams (1963) originally used a host differential to distinguish six *A. candida* races including race 1 - *Raphanus sativus*, race 2 – *B. juncea*, race 3 - *Armoracia rusticana*, race 4 - *Capsella bursa-pastoris*, race 5 - *Sisymbrium officinale* and race 6 - *Rorippa islandica*. Verma *et al* (1975) then described race 7 from *B. rapa*. Race 8 was documented by Delwiche and Williams (1977) on *B. nigra*. Race 9 is specific to *B. oleracea,* race 10 to *Sinapis arvensis* (Hill *et al*., 1988) and race 11 is adapted *B. carinata* (Williams, 1985). Races 12 – 17 were described based on a host differential of different Indian varieties of *B. rapa* and *B. juncea* (Verma *et al*., 1999, Gupta & Saharan, 2002). *A. candida* can also specialise amongst varieties of the same species, with sub-races described for race 2 and race 7 based on virulence (V) or avirulence (A) on different cultivars of *B. rapa* and *B. juncea*. Pathotypes are considered to be most virulent on the plant species from which they were isolated (homologous host) whilst also able to infect certain genotypes of closely related *Brassica* species, particularly where hosts share significant genetics (Rimmer *et al*., 2000).

The wide range of host adaptation and race specialisation of *A. candida*, which represents a single species, is of wider interest to research of host-pathogen evolution. Most plant pathogens have restricted host ranges due to host specialization and cannot exploit even closely related hosts. This is because adaptation of pathogen effectors that facilitate access to one host may subsequently impede access to others by triggering *R*-gene mediated innate immunity, which subsequently drives host-specialisation. *A. candida* seems to have overcome these issues however, having evolved multiple host-specialised races or pathotypes.

The first major insights into *A. candida* race evolution have been facilitated by advances in next-generation-sequencing (NGS) technology and publishing of the reference genome in 2011 (Links *et al*., 2011). Four years later, a study by (McMullan *et al*., 2015) produced complete genome sequence for several isolates to reveal three genetically divergent (~1%) host-races, each showing evidence of historical recombination. In some rare instances, recombination between *A. candida* pathotypes is thought to generate novel effector repertoires that enable adaptation to a new host, which is proceeded by rapid asexual propagation of the pathogen. Understanding of *A.*

candida race evolution has since been advanced by (Jouet *et al*., 2019) who adopted pathogen enrichment sequencing (PenSeq) to show that each host plant species supports colonisation by one of 17 distinct phylogenetic lineages, each containing a unique arsenal of effector alleles.

Albugo spp. have evolved the ability to supress host innate immunity, thereby enabling secondary infection by other typically avirulent pathogens (Cooper *et al*., 2008). Subsequent colonisation of the host by multiple physiological races could then enable sexual recombination and genetic exchange between races to produce variants that could colonise new hosts (Hedrick, 2013). It has now been demonstrated *in vivo* using sequential inoculations that infection by a virulent race can facilitate colonization by a second otherwise avirulent race to potentially enable sexual reproduction between pathotypes (McMullan *et al*., 2015). In many cases, sexual recombination between different pathogen races could put them at an evolutionary disadvantage, with novel combinations of inherited effectors putting them at high risk of detection by immune systems of both host species. However, contemporary monoculture cropping systems exert extreme selection pressure on pathogens, where once a variant with basic compatibility is produced it can proliferate rapidly throughout crops.

Enhanced susceptibility imposed by *A. candida* may also enable colonisation and adaptation of other pathogen species such as *Hyaloperonospora parasitica* to an *Albugo*-susceptible host (Thines, 2014). This appears to be a maladaptive trade-off against the benefits conferred by sexual reproduction, as it leads to intraspecific competition for access of the same resources (Cooper *et al*., 2008).

1.6 White Rust Resistance

The most extensive work on the genetics of white rust resistance (WRR) have been undertaken using the *A. candida – A. thaliana* pathosystem (Borhan *et al*., 2008; Borhan *et al*., 2004; Borhan *et al*., 2010; Cevik *et al*., 2019). Early research efforts focused on identifying resistance in *A. thaliana* to race 4 isolates and were able to identify three resistance loci (*RAC1*, *RAC2* and *RAC3*) each conferring phenotypes of reduced blister formation or complete lack of asexual reproduction by the parasite. The *RAC1* locus was fine mapped in resistant accession Ksk-1 to chromosome 1, between restriction fragment length polymorphism (RFLP) markers m253 and m254 (Holub *et al*., 1995). Positional

cloning was later used to identify the dominant *RAC1* gene as a TNL *R*-allele, which was found to be dependent on a functional copy of the lipase-like defence regulator *EDS1* (Borhan *et al*., 2004). Use of recombinant inbred mapping located the recessive *RAC2* loci in Ksk-2 to a 6 cM interval on the bottom arm of chromosome 3, whilst *RAC3* (Ksk-1) was mapped in close linkage to the *RPP8/HRT* resistance complex on chromosome 5 (Borhan *et al*., 2004).

Importantly, several *A. thaliana* accessions were later discovered that permitted varying degrees of infection by *A. candida Brassica* crop races 2, 4, and 7, enabling study of non-host resistance (NHR) mechanisms for these agriculturally important pathotypes. The white rust resistance gene *WRR4* was identified as an *EDS1* dependent TNL allele in the accession Col-0, located on chromosome 1 that confers full immunity to each race (Borhan *et al*., 2008; Borhan *et al*., 2010). It has since been determined that Col-0 contains three other WRR alleles (*WRR5*, *WRR6* and *WRR7*) situated on chromosome 5 (Holub & Cevik, unpublished). A recent study by Cevik *et al.* (2019) of *A. thaliana* NHR to races 2 and 9 was able to produce two susceptible transgressive segregant lines from a MAGIC inbred line population which were crossed back to each MAGIC parent for mapping in the F_2. It was found that resistance to the race 2 isolate Ac2v could be explained by up to four genes that include two adjacent paralogues of *WRR4* (*WRR4A* + *WRR4B*), *WRR8* (chromosome 5) and *WRR9* (chromosome 1). Whilst resistance to the race 9 isolate AcBoT was conferred by *WRR12* on chromosome 1. All these genes belong to the TNL class of *R*-genes. Three *A. candida* isolates have been collected from different *Arabidopsis* species that are virulent in Col-0 (overcoming *WRR4* resistance), and one of these isolates, Ac-Exeter, was used to map an alternative *WRR4* allele in *A. thaliana* accession Oy-0 (Fairhead, 2016).

Assessment of WRR in the major *Brassica* crop species has been met with mixed results, with several studies focusing on preliminary identification of resistant sources of germplasm. *B. oleracea* for example remained largely unexplored for WRR until (Santos & Dias, 2004) screened a core collection of 400 accessions for disease expression using a Portuguese race 2 isolate Ac502. They identified 47 sources of resistance, with nine accessions presenting sufficient disease index scores for consideration in breeding programs. An important source of recessive resistance has recently been found in the doubled haploid *B. oleracea* accession EBH527, which provides broad spectrum resistance

when tested with 18 race 2 isolates at the Wellesbourne Crop Centre, UK. The resistance has been fine-mapped to an 18 kb interval on chromosome 2 (ACA2), that contains four candidate genes (Fairhead, 2016). Further work to identify the functional gene will be a focus of Chapter 3.

Inheritance studies in *B. napus* have found sources of dominant resistance in accession BN-Sel where progeny derived from an F_1 backcross suggest control by dominant duplicate genes (Verma & Bhowmik, 1989). A study by Fan *et al*. (1983) made two independent crosses between the race 7 resistant Canadian cultivar 'Regent' and two susceptible Chinese lines to identify independent dominant genes at three loci designated *Ac7-1*, *Ac7-2*, and *Ac7-3*. In *B. juncea* WRR has been shown to be inferred by monogenic dominance (Ebrahimi *et al*., 1976; Tiwari *et al.,* 1988; Paladhi *et al*., 1993; Somers *et al.,* 2002; Vignesh *et al.,* 2009; Panjabi-Massand *et al.,* 2009; Vignesh *et al.,* 2011).

Earlier studies such as (Cheung *et al*., 1998) used densely populated RFLP markers to map a 6.7 cM interval (*Ac2*) on chromosome 7 in the resistant accession J90-2733. Subsequently, (Varshney *et al*., 2004) developed co-dominant and more precise PCR-based cleaved amplified polymorphic (CAPS) markers in closer proximity to the Ac2 locus. Later studies used a variety of marker types including intron polymorphic (IP) markers, amplified fragment length polymorphism (AFLP) markers and simple-sequence repeat (SSR) markers to map three previously established WRR loci (Panjabi-Massand *et al*., 2010), which were then followed up by marker validation in known resistant cultivars (Singh *et al*., 2015). Two independent resistances have since been mapped in Eastern European *B. juncea* lines, with partial resistance in Heera identified on linkage group A4 (AcB1-A4.1) and complete resistance in Donskaja located on linkage group A5 (AcB1-A5) (Panjabi-Massand *et al*., 2010). The causal resistance gene was recently determined as a canonical CNL protein using a combination of positional cloning and a candidate gene methods, and verified as the resistance gene via transformation into the susceptible *B. juncea* background Varuna (Arora *et al*., 2019). This *R*-gene called *BjuWRR1* confers broad spectrum resistance against a collection of six race 2 *A. candida* isolates collected from across India (including Cutlass virulent AcB1).

In *B. rapa*, a study by Santos *et al*., (2006) screened for WRR in several genebanks before studying inheritance patterns of the most resistant pak choi line BRA 117 by crossing to the rapid cycling line CrGC 1.19. Analysis in the F_2 segregants suggested that

resistance is controlled by two nuclear genes with a dominant recessive interaction. Another study by Kole *et al*., (2002) utilised 144 RFLP markers to map WRR using a race 2 and a race 7 isolate from a biparental cross between resistant accession Per and susceptible R500. A single, major effect locus (Aca1) was identified on linkage group A04 for both races, suggesting either control of both isolates by a single gene, or two tightly linked genes. A second minor effect QTL was also discovered on A02, which is syntenic to the race 9 *B. oleracea* locus studied here in chapter 3.

1.7 Plant genebanks

Since Vavilov created the first *ex situ* genebank in the 1920s (The Institute of Plant Industry, Petersburg) they have since become a key global resource for long term storage and conservation of plant genetic resources (PGRs), where material is made available to plant breeders, researchers and general users for crop development. Today, there are about 1700 gene banks globally, storing an estimated 7.4 million accessions; comprising of advanced cultivars, landraces and wild type varieties (On *et al*., 2019). Approximately 50% of global gene bank accessions are made up of the top ten agriculturally important species, where 28% of which only represent wheat, rice and barley (Kilian & Graner, 2012). However, minor crops, crop wild relatives and under-utilised crops are still widely under-represented. This is problematic as crop wild relatives and locally adapted landraces contain rich allelic repertoires, most of which have been selected out of elite crop varieties over centuries of intensive breeding. Stored PGR diversity therefore represents an invaluable resource for future breeding strategies that are able to cope with challenges of sustainable agriculture and food security.

Whilst the proliferation of global genebanks over the past century is testament to the progress made in germplasm conservation, the resources they contain are still widely under-exploited. A study by (Sharma *et al*., 2013) for example, estimated that as little as < 1% of accessions conserved have ever been used in crop development. The large sizes of genebank collections are inherently part of the problem, where characterisation, evaluation, utilisation and maintenance of such large sets of broadly distributed material is inherently difficult (Odong *et al*., 2013). For example, passport and genotype data suggests that < 30 % accessions across global genebanks are unique accessions and not duplicated (On *et al*., 2019).

For plant breeders, selecting the appropriate accessions to include in working collections can be problematic, where detailed information on phenotype is often lacking. Core collections, an idea first proposed by Frankel (Frankel & Brown, 1984) go some way to address these issues where smaller subsets of germplasm are selected that maximise the range of genetic diversity represented within the overall collection, thus making their effective management and application much simpler. However, useful alleles often miss out from inclusion in core collections, particularly for traits such as disease resistance where *R*-alleles are frequently rare within or amongst populations. (Reyes-Valdés *et al*., 2018).

1.8 Advances in molecular mapping technology

Advances in next-generation sequencing technologies (NGS) are beginning to unlock the potential of germplasm resources and are rapidly becoming an integral part of genebank management (Kilian & Graner, 2012). The continuously falling prices of NGS have widened availability of the technology within research, opening the door to production of thousands of genetic markers for characterising biological variation. These vast datasets allow for precise classification of stored accessions in genebanks, aiding in studies of relatedness as well as management of duplicated material. For marker-trait association studies, the high marker density provided by NGS allows trait variation to be mapped with high precision, facilitating the discovery of candidate genes underlying key agronomic traits. This is all enhanced by the continuous publishing of new and improved plant reference genomes, providing overdue research on non-major crop varieties. Affordability of NGS is enhanced further by improvements in DNA barcoding and sample multiplexing, allowing simultaneous high-throughput sequencing of hundreds of individuals (Buermans & den Dunnen, 2014).

Accessing genetic diversity to identify genes controlling traits of interest is therefore no longer constrained by the availability of genetic information. However, technicalities regarding the mapping processes themselves place an inherent limitation on the sampling of germplasm diversity. For example, traditional linkage mapping methods, (details of which are reviewed by Collard *et al*. (2005)), are frequently reliant on producing mapping populations from bi-parental crosses, where diversity is restricted to that of two genetically fixed parents. This represents a tiny amount of the variation

contained within *ex situ* genebanks. The method is labour and resource intensive, requiring assessment of hundreds of individuals for generating phenotype and genotype data. Subsequently only a very small number of mapping populations are ever advanced at once, which effectively bottlenecks the screening and discovery of useful alleles for crop development.

Association mapping, which relies on linkage disequilibrium to assess the relationship between phenotypic variation and genotype (reviewed Breseghello & Sorrells (2006)) goes some way to correcting for these issues. Instead of producing bi-parental crosses, collections of advanced cultivars landraces or wild species can be utilized directly to identify marker-trait associations. Benefits of this include: 1) access to a greater diversity of allelic variation; 2) a greater likelihood of higher resolution mapping due to leveraging historical recombination events; 3) no need for production of bi-parental mapping populations, saving considerable time, effort and expense. This method, typically conducted in the form of genome-wide association studies (GWAS), requires the utilisation of hundreds of individuals to raise both the numbers of alleles within the study and the historical recombination events which add power to the analysis.

Bulked Segregant Analysis (BSA) is now a widely used strategy for rapidly identifying markers across the genome that associate with a trait of interest. The method was first developed in the early 1990s (Michelmore *et al*., 1991) and resides on the idea that a bulked sample containing a group of individuals with the same phenotype will contain all the alleles for the trait. Therefore, where two bulked pools of individuals that differ for a particular trait are compared, the allele frequency will only significantly deviate at the locus controlling this trait. The major benefit of BSA over both linkage mapping and association mapping studies is that it does not require genotyping of hundreds of individuals within a segregating population. Instead, individual plants are grouped into two bulked samples according to whether they represent an extreme level of the particular trait. With only two DNA samples then required for analysis, this provides substantial savings in time and costs. Importantly, the savings made through use of BSA allow for a greater amount of germplasm to be processed and assessed, thereby enhancing the use of genebank diversity for marker-trait discovery.

BSA can be utilised with any co-dominant marker dataset and has subsequently been adapted for next-generation sequencing (NGS) to maximise provision of SNPs for

analysis (NGS-BSA). Within a segregating F_2 population, SNPs unlinked to the trait of interest are expected within approximately 50% of the reads, whilst SNPs linked to the trait should be over- or under- represented in comparison (closer to 0 or 100%). Computational methods such as QTL-seq developed by Takagi *et al.*, (2013) generate SNP-indices (defined as the number of reads containing a SNP divided by the total sequencing depth at that SNP) to enable rapid detection of quantitative trait loci (QTLs).

1.9 Research objectives

Chapter 2.

Developing novel mapping methods to rapidly identify major effect disease resistance loci directly from genebank accessions.

Objective 1. Developing a Genotyping-by-Sequencing (GBS) method for mapping WRR loci directly from a genebank diversity collection in *Brassica rapa*.

Objective 2. Application of whole genome resequencing (WGS) and bulked-segregant-analysis (BSA), to map WRR QTL directly from genebank *Brassica rapa* collections.

Chapter 3.

Investigating broad-spectrum white rust resistance in *Brassica oleracea* accession EBH527.

Objective 1. Molecular investigation of a predicted non-NLR determinant of white rust resistance in *B. oleracea*.

Objective 2. Mapping a dominant white rust resistance in *Brassica oleracea* using WGS-BSA.

Chapter 4.

Pathogenomics of *Albugo candida* race 2 isolates from English mustard production in the UK.

Chapter 2

Developing a novel method to map and identify major effect disease resistance genes directly from genebank accessions

2.1 INTRODUCTION

Brassica juncea (oilseed mustard) is an important multi-purpose crop for the economy and health of people living throughout south Asia, and is the *Brassica* used in English mustard. In India for example, *B. juncea* is used for production of cooking oil from seeds, green vegetable and animal fodder from leaves, and domestic cooking fuel from dried straw after harvest (Shekhawat *et al*., 2012). India is among the major rapeseed-mustard growing countries in the world and ranks second in area and third in production after China and Canada. Indian mustard comprises more than one third of the home-grown vegetable oil production in India and covers ca. 7.0 million hectares. The crop is cultivated mostly by resource poor farmers in smallholdings (1 hectare or less). Annual yields from this production are around 6.4 million tonnes and a local market value of about £ 2.4 billion. Similar production and consumption of *B. juncea* occurs elsewhere in south Asia and China. The production per hectare of oilseed mustard in India is lower than the world's average due to a large number of factors including cultivation under marginal (low rainfall) conditions, and losses caused by various biotic and abiotic stresses.

Albugo candida (*Brassica* white rust) is the major biotic constraint of *B. juncea* oilseed production in the Indian subcontinent, where it can cause yield losses of 30-60% (Li *et al*., 2009). The most severe damage results from 'staghead' formation, where impaired floral development results in pod abortion and loss of seed (Saharan *et al*., 2014). Almost all the major varieties of *B. juncea* belonging to the Indian gene pool are highly susceptible to white rust. Whilst fungicides such as metalaxyl and garlic extract can provide partial control of the disease (Gairola & Tewari, 2019), the use of chemicals is ecologically undesirable. Furthermore, the costs inferred from their usage adversely affect small and marginal farmers. Breeding for white rust resistance (WRR) is therefore the most efficient, cost-effective, and environmentally sustainable way of protecting *Brassica* crops from white rust disease. Due to the potential for rapid evolution in *A. candida*, it is important that as many sources of WRR genes are identified as possible that when combined by pyramiding using conventional breeding methods, by transgenic 'stacking' of resistance alleles in a single construct using recombinant DNA methods, or by mutational knock-out of dominant susceptibility alleles using gene-editing methods (see Chapter 3) will provide the best chance for durable disease control.

White rust resistance is considered to be relatively rare in germplasm collections of *B. juncea* (Anupriya *et al*., 2020; Awasthi *et al*., 2012; Panjabi-Massand *et al*., 2010). This was also observed in screening of a *B. juncea* collection from the UK Vegetable genebank (Wellesbourne, UK) with two UK isolates of *A. candida* race 2 where < 10% of accessions contained resistance (Holub & Vicente, unpublished). However, Indian researchers identified two Eastern European accession with white rust resistance (WRR) including one Donskaja which appeared to confer broad spectrum resistance to a collection of six Indian isolates (Panjabi-Massand *et al*., 2010). The causal resistance allele (designated *BjuWRR1*-Donskaja) was recently determined as a canonical CNL protein using map-based cloning, and verified via transformation as the dominant resistance allele in the susceptible accession Varuna (Arora *et al*., 2019). Subsequently, Dev *et al*. (2020) characterised a larger collection of Indian *A. candida* race 2 isolates and identified four that were virulent in Donskaja (Ac-BjBio-Pant, Ac-BjV-Pant, Ac-Bna-Pant and Ac-Alw). Similarly, Donskaja is a host differential between two isolates of the pathogen (AcBj12-avirulent and AcBjDC-virulent) that were collected from an English mustard farm. Thus, durable white rust control in *B. juncea* will require a combination of at least two complementary *WRR* alleles.

In contrast, *Brassica rapa* is an excellent resource as a secondary gene pool for Indian mustard of additional WRR alleles. In a European Union genetic resources project (RESGEN CT99 109-112; https://ecpgr.cgn.wur.nl/Brasedb/brasresgen.htm), 95 out of 100 accessions in a diversity collection of *B. rapa* were reported as heterogenous populations of resistant and susceptible individuals following inoculation with a Canadian isolate of *A. candida* race 7 (naturally occurring pathotype in *B. rapa*). A subset of fifteen accessions from this *B. rapa* collection all contained resistant individuals following inoculation with AcBj12, a UK isolate of *A. candida* race 2 from *B. juncea* (Holub EB and J Brough, unpublished). These accessions represented a global distribution and range of vegetable and oilseed crop types of *B. rapa.* Although races 2 and 7 can be outcrossed in the laboratory (Adhikari *et al*., 2003) they are substantially divergent within the species (Jouet *et al*., 2019) so the high occurrence of resistance in *B. rapa* to the *B. juncea* pathotype was unsurprising.

The aim of this chapter was to determine whether genes conferring major effect white rust resistance could be mapped and identified directly from genebank accessions.

Given the extensive resource of potentially novel resistance alleles, conventional linkage mapping (LM) by selecting individual resistant plants from each accession and cross-pollinating with a susceptible control genotype would provide a methodical but slow and resource intensive process. Alternatively, a bulked segregant mapping approach could take advantage of that fact that *B. rapa* is an outcrossing species, and that genebank accessions have been maintained from collection of advanced cultivars, landraces and wild species as small intra-pollinated populations (referred to henceforth as accessions). It should therefore be possible to apply mapping to rapidly identify WRR loci directly from the phenotypically heterogeneous genebank accessions.

Thus, the specific objectives of this chapter were to test two DNA-sequencing methods for generating genome-wide genotyping data. This included: Objective 1, testing Genotyping-by-Sequencing (GBS-mapping) which potentially offers a cost-effective way to genotype individual plants; and Objective 2, Whole Genome Re-sequencing (WGS) of pooled DNA which could potentially combine with conventional bulked segregant analysis (WGS-BSA).

2.2 MATERIALS AND METHODS

2.2.1 Maintenance of *Albugo candida* isolates. The UK *A. candida* race 2 isolates AcBj12 and AcBjDC were collected in 2016 by Professor Eric Holub's group from an English mustard farm near Thorney (east of Peterborough). To generate genetically uniform cultures, susceptible *B. juncea* cv Sutton were inoculated with zoosporangia harvested from single small pustules, using a pipette tip. This was suspended in a small volume of sterile water before being applied to the underside of nine-day-old cotyledons. After two cycles of inoculations from single pustules, the two refined isolates were then used to generate bulked sporangial inoculum on a tray of Sutton seedlings for long term storage in -80 freezers. The race 2 isolate Ac2v was recovered from Wellesbourne freezer collections on *B. juncea* cv Sutton and was maintained and disposed of under the enhanced quarantined conditions required by a licenced isolate. Whilst the race 9 isolate AcBoWells (used in Chapter 3) was collected from a cabbage field experiment at the University of Warwick Crop Centre and maintained on *B. oleracea var* Maris Kestrel.

To prepare plants used for maintaining isolates, P180 plug trays (2.5 x 2.5 cm cells), cut to 10 x 6 grids were filled with Levington M2 compost, and were left to soak for 30 minutes in a tray of water. Holes of 0.5 cm depth were made in the compost for each cell, and a single seed was sown before covering with a fine layer of vermiculite. Plug trays were then placed in propagators with a clear sealed lid and stored in a Conviron plant growth cabinet at 20±2°C with a 10h photoperiod.

To quantify inoculum, the concentration of suspended zoosporangia was assessed using a haemocytometer and adjusted to approximately 4 x 10^4 zoosporangia per ml. Plants were inoculated using two 10 ul droplets applied to each half of the upper cotyledon surface with a pipette. Propagators were then sealed and placed back into the Conviron at 20±2°C for an initial 24h period of darkness, followed by a 10h photoperiod. Depending on the isolate, pustules would emerge 7-10 days-post-infection (dpi) and were subcultured onto fresh plants every two weeks.

2.2.2 Selection of *Brassica rapa* genebank accessions. Seed was obtained from the UK Vegetable Genebank in Wellesbourne. These were accessions which Prof Holub had previously tested using a Canadian isolate of *A. candida* race 7 (Ac7v) and AcBj12. Approximately 50 plants were sown per line and inoculated at the cotyledon stage, as described in maintenance of *A. candida* isolates. Eight lines which exhibited heterogeneity for resistance and susceptibility were chosen for subsequent mapping experiments (Table 2.1). These accessions, which originate from between East Europe and East Asia represent a mixture of commercially viable varieties, landraces and wild relatives. Collectively they present a range of morphological and varietal types (Figure 2.6).

2.2.3 Inoculation of experiments, and tissue sampling. The eight experimental lines of *B. rapa* were sown, inoculated and maintained in the same manner as described in maintenance of *A. candida* isolates. These were divided between two trays per line (50 plants in each), sown with two susceptible control cultivars Burgonde and Cutlass (five plants of each), with no resistant control available. The experiment was set up and conducted twice using both UK race 2 isolates AcBj12, and AcBjDC.

After 10 days-post-inoculation (dpi) plants were assessed for disease symptoms on both the upper and lower cotyledon surface, and given quantitative scores based on the eight-class scale of phenotypic variation described by (Fairhead, 2016; Figure 2.4). Plant tissue was harvested from the first primary leaves of phenotyped seedlings using a sterile 6mm Harris Uni-Core Leaf Punch. This was rinsed in 70% ethanol and wiped dry between sampling of each plant. For each line, a fully resistant (no sporulation; a phenotype score ≤ 1), and fully susceptible (unrestricted pustule formation within 7-10 dpi; a phenotype score ≥ 4) were produced, as well as individual samples of every plant contained in the bulks. Five leaf disks were taken per plant, with a single disk contributing to the relevant phenotype bulk and the remaining four used to produce individual plant samples stored in 2 ml Eppendorf tubes. The number of individuals per bulk was determined by the ratio of distinct resistant vs susceptible individuals available across the two experimental trays (Figure 2.4). Tissue samples were then freeze dried for 48 hours and stored at -20 C.

2.2.4 DNA extraction. Freeze dried tissue samples were homogenised by adding two sterile tungsten steel beads to each Eppendorf tube which were then sealed with parafilm and placed into the TissueRuptor® for a total time of 1.5 m at 60 Hz. The two TissueRuptor® sample plates were removed after 45 s and inverted to ensure samples received equal levels of reverberation and subsequent tissue disruption. Samples were visually inspected to ensure each was completely homogenised to powder before being stored on ice.

DNA extraction was undertaken using a modified CTAB protocol from (Afanador *et al*., 1993), with volumes of reagents doubled for bulk samples to account for the increased tissue mass. Chemical lysing of cell membranes was performed by adding 400 ul of CTAB (Sigma Aldrich) with 5ul of RNase A to each sample. These were vortexed, spun down and incubated in a water bath for 1 hr at 60 °C, agitating samples every 20 minutes. The lysate was then centrifuged for 15 minutes at 13,000 rpm before careful removal and transfer of the supernatant (without disturbing the tissue layer) to fresh 1.5 ml DNA LoBind® Eppendorf tubes. To separate the aqueous phase (containing the DNA) from the organic phase, 400 ul of phenol:chloroform:isoamyl alcohol (25:24:1) was added and samples were inverted 30 times to produce an emulsion. This was centrifuged for 15

minutes at 13,000 rpm and the supernatant aqueous layer was carefully removed and transferred to fresh 1.5 ml DNA LoBind® Eppendorf tubes. To remove residual lipids and organic matter, this last step was repeated a second time from the addition of phenol:chloroform:isoamyl alcohol (25:24:1). DNA was then precipitated out of solution with the addition of 280 ul of ice-cold isopropanol (0.7 x volume), inverting tubes 30 times before storing them on ice for 5 minutes. Samples were centrifuged at 13,000 rpm for 10 minutes to pellet the DNA before discarding the isopopropanol. To remove residual salts, the pellet was then washed twice, first with 70% ethanol and then 90% ethanol. After carefully discarding the 90% ethanol (using a pipette so as not to lose the pellet) samples were air dried for 10 minutes on the lab bench before eluting in 50 ul of TE buffer. To ensure complete re-suspension of DNA, samples were incubated in a water bath at 50 ° C for 30 minutes. From the final eluted stocks, aliquots were prepared as 1:10 dilutions to provide DNA template for downstream PCRs.

DNA yields were quantified using a Qubit dsDNA BR Assay Kit with a Qubit 2.0 Fluorometer (Invitrogen, Waltham, USA) to confirm adequate concentrations of ≥ 50 ng μl−1 were achieved. DNA quality was assessed using a NanoDrop Spectrophotometer (NanoDrop, Wilmington, USA). Electrophoresis was conducted on 1% agarose gels, stained with GelRed (Biotium, Fremont, USA), and visualised using an ultraviolet transilluminator to ensure extracted DNA was of high molecular weight with minimal shearing.

2.2.5 Genotyping-by-Sequencing (GBS). For GBS, extracted DNA stocks were eluted to provide 3 ug per sample, at an approximate concentration of 50 ng/ul. This included 207 samples in total, split across eight accessions. Independent resistant samples were prepared for sequencing, with the number sent determined by the availability of phenotypically distinct resistant plants available for each line (Table 2.1 and Figure 2.4). For the susceptible phenotype, DNA from two biological replicate tissue bulks were used per line. The logic here was that sources of WRR in these *B. rapa* lines would more likely be dominant than recessive, due to expected heterozygosity at the causal locus in resistant individuals. Maintaining the resistant samples as individuals, rather than a single bulk, would provide clear resolution of any such heterozygous states in downstream analysis. Samples were sent to LGC Genomics facility in Berlin, Germany, where

sequencing libraries were prepared through restriction digests at *Ape*KI recognition sites, using the methods described by (Elshire *et al*., 2011). Samples were each ligated with index barcodes and multiplexed before running on an Illumina NextSeq 500 V2, to produce single-end reads (75 bp) using 96-plex sequencing (150 cycles).

2.2.6 Whole genome resequencing (WGS) of DNA bulks. For WGS, five *B. rapa* genebank accessions (EH_5, EH_25, EH_47, EH_61 and EH_95) were chosen and each used to generate two phenotype DNA bulks (one resistant and one susceptible) for further mapping of WRR loci using a method described by Mansfeld and Grumet (2018). Selection of these accessions was based on a combination of factors including that they produce distinct resistance phenotypes and represent a wide geographical range with variable genetic structure (Figure 2.9). To provide a comparative assessment of mapping with Donskaja-virulent isolate AcBjDC, resistant and susceptible DNA bulks previously generated from line EH_25 (tested with AcBjDC) were also included for sequencing. Factors that favoured EH_25 for testing with a second isolate included: 1) this accession is represented by the highest number of resistant individuals for GBS-mapping, and thus demonstrates the strongest association to phenotype in these results (Objective 1); 2) resistant individuals exhibited a distinct host response following inoculation with either AcBj12 and AcBjDC and therefore provided a bulk sample that minimises risk of false negatives (susceptible individuals that escaped infection); 3) as a wild species (not a crop type) results can inform about the possibility of mapping in a non-domesticated accession.

Resistant and susceptible DNA bulks for WGS were produced from the same sample extractions that were used for GBS. For a resistant bulk this required combining equal amounts of individual DNA stocks, as determined by a Qubit 2.0 Fluorometer. Final bulks were eluted to provide a total of 3.5 ug at a concentration > 50 ng/ul per sample. The numbers of individuals represented in each bulk is displayed in Table 2.1. Samples were submitted to Novogene (UK) in Cambridge, for library construction and sequencing. Sequencing libraries were generated using a NEBNext® DNA Library Prep Kit (following manufacturers recommendations) and indices were added to each sample. DNA was fragmented to a size of 350 bp using sonication, with fragments end polished, A-tailed and ligated with the NEBNext adapter for Illumina sequencing. Finally, libraries were PCR enriched, purified (AMPure XP system) and quantified using qPCR. Paired-end (150 bp)

sequencing was performed on an Illumina Novaseq 6000. To account for heterozygosity in the genebank accessions, WGS (Novogene Ltd) was performed at an enhanced sequencing depth of 90x.

2.2.7 Bioinformatic pipeline of GBS and WGS-BSA datasets. Bioinformatic support for advancing next-generation sequencing datasets was provided by Richard Stark and Dr Laura Baxter. Raw GBS reads were processed in-house by LGC to produce a single reference-based SNP dataframe containing genotype information for all 207 samples. This began with demultiplexing of raw reads based on ligated sample barcodes, which were then adapter clipped and removed. Reads where 5' ends did not match an ApeKI restriction site were also removed. Further quality trimming of adapter clipped reads included: 1) removal of reads containing Ns; 2) Trimming of reads to maintain a minimum average Phred quality score > 20 (equating to a 1 in 100 probability of an incorrect base call); 3) Removal of reads with final length < 20 bp. All read quality reports were produced using FastQC (Andrews, 2010).

The GBS alignment was performed against the *Brassica rapa* IVFCAASv3 reference assembly (Wang *et al.*, 2011) using the quality trimmed reads and the Burrows-Wheeler Aligner (BWA) (v0.7.12, (Li & Durbin, 2009)). For variant discovery and genotyping Freebayes (v1.0.2-16, (Garrison & Marth, 2012)) was applied with parameter settings including; a minimum base quality of ten, a read mis-match limit of five, and a minimum coverage of five. Initial filtering of variants was then performed using a GBS-specific rule set which determines that read count for a locus must exceed eight reads and contain a minimum allele frequency (MAF) across all samples of 10%.

For WGS, raw reads were received from Novogene that had undergone initial filtering for removal of low-quality reads, such as where N > 10% (with N representing an undetermined base) and reads where low-quality bases (Qscore <= 5) represent > 50% of the total bases. Reads containing sequencing adaptors were also removed, and quality control reports were checked in FastQC (v0.11.9). From this point the data was processed through the GATK germline short variant discovery (SNPs + Indels) pipeline (Figure 2.1). Filtered paired-end reads were aligned to the *Brassica rapa* IVFCAASv3 reference, using BWA (v0.7.12) with default settings. The aligned *.sam files were then soft-clipped using CleanSam (Picard v2.23.7, GitHub Repository. http://broadinstitute.github.io/picard/;

Broad Institute), before being converted to sorted ∗.bam files using Samtools (v1.9; Li *et al*. (2009)). Any duplicate reads were then marked for downstream analysis using MarkDuplicates (Picard v2.23.7). Final error checks of ∗.bam files were then made using the Picard tool ValidateSam to ensure files were ready for calling SNPs and INDELs. This was performed using HaplotypeCaller (GATK tool, Poplin *et al*. (2017)), which performs *de novo* assembly of any variable region to accurately and simultaneously call SNPs and INDELs. Outputted gVCF files for both phenotype bulks were consolidated per line into a single gVCF (GATK tool, GenotypeGVCFs) and converted to Table format (GATK tool, VariantsToTable) for marker-trait analysis.

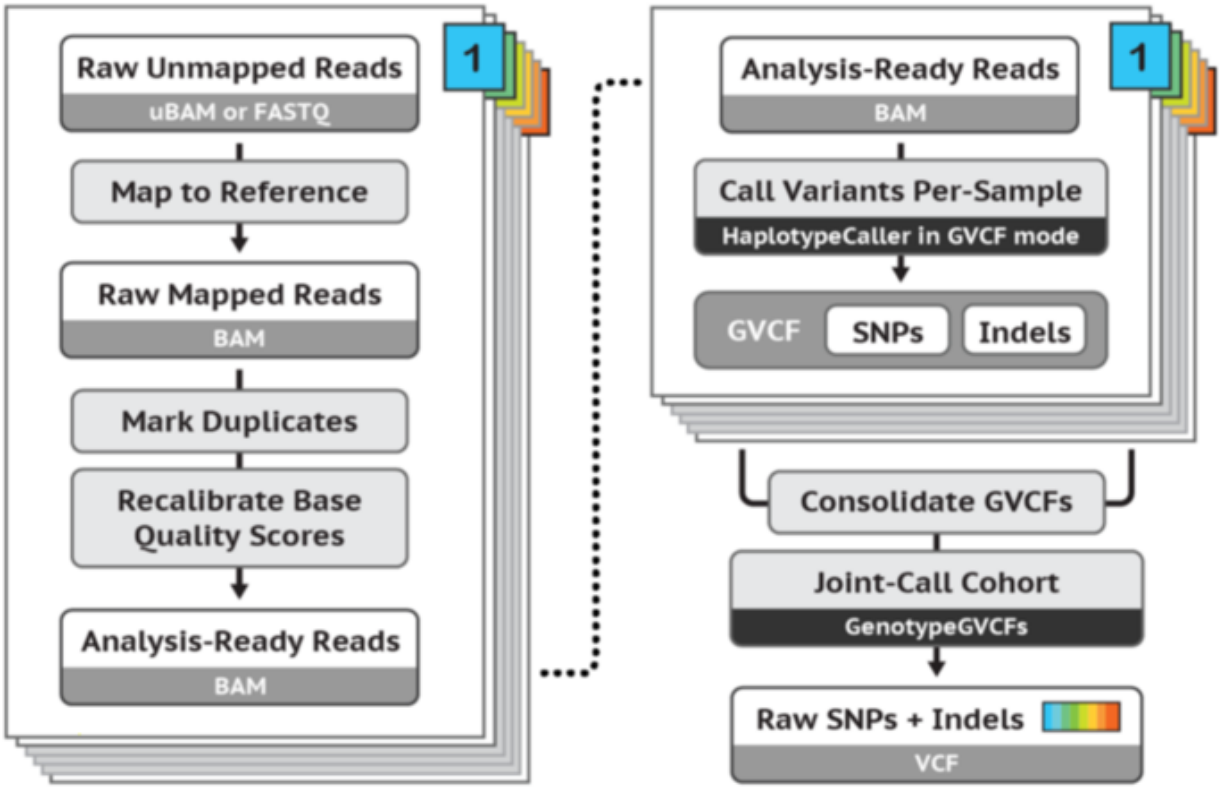


Figure 2.1. The GATK (Genome Analysis Toolkit) germline short variant discovery (SNPs + Indels) pipeline, utilized here for advancing *Brassica rapa* phenotype bulked WGS data through to consolidated and genotyped (SNP + INDEL) gVCF files.

2.2.8 Genome-wide mapping analyses. For the GBS datasets, the SNP variant dataframe containing filtered genotypes for all 207 samples (eight accessions) was processed for identification of WRR associated SNPs (raSNPs) using a combination of Excel® (v16.43) and R (v1.1463) based platforms. The dataframe was initially processed for: 1) removal of rows where > 5% of sample genotypes were missing data; 2) removal of rows where the two biological replicate susceptible bulks display different genotypes; 3) removal of SNPs aligned to scaffolds; 4) for each row, counting the number of resistant individual genotypes that matched with the susceptible bulk, termed MatchCount; and 5) retaining rows where MatchCount is ≤ 5%. From the remaining raSNPs, filtering was performed to identify both dominant and recessive raSNPs by selecting rows where the susceptible bulk was either homozygous (dominant) or heterozygous (recessive) respectively.

For the WGS-BSA datasets, QTL analysis was performed using WGS datasets and QTLseqr (v0.7.5.2, (Mansfeld & Grumet, 2018)), which is an R package currently available for BSA-NGS mapping. The G' statistical approach (Magwene *et al*., 2011) was applied, which uses a tricube smoothed G statistic to identify and assess statistical significance of QTL.

2.2.9 *R*-gene annotation.

Disease Resistance Analysis and Gene Orthology (DRAGO 2) software (Osuna-Cruz *et al*., 2018) was used here to annotate the *B. rapa* IVFCAASv3 reference assembly and visualise the position of identified raSNPs relative to *R*-genes. The tool applies 60 profile hidden Markov modules (HMM) from the HMMER v3 package (Finn *et al*., 2011) to automatically annotate and predict *R*-genes from DNA or amino acid.

2.3 RESULTS

2.3.1 Characterising virulence phenotypes of three race 2 *Albugo candida* isolates on *Brassica juncea* accession Donskaja. The two UK isolates AcBj12 and AcBjDC, as well as Canadian isolate Ac2cv were inoculated onto Donskaja cotyledons for an assessment of virulence phenotype regarding *BjuWRR1* resistance. Sporulation was observed for all three isolates ten days post inoculation, with minor levels present of AcBj12, moderate levels with AcBjDC and major levels with Ac2v. For clarity, as delayed, minor sporulation indicates elicitation of a host response, AcBj12 has been categorised here as Donskaja-avirulent, whilst AcBjDC and Ac2v are described as Donskaja-virulent. These virulent isolates are therefore considered as *BjuWRR1* resistance-breakers, with no elicitation of a host response. Of note, confirmation of AcBjDC and Ac2v as Donskaja-virulent was not made until the latter part of this project, which is why these isolates were not prioritised here for mapping of WRR.

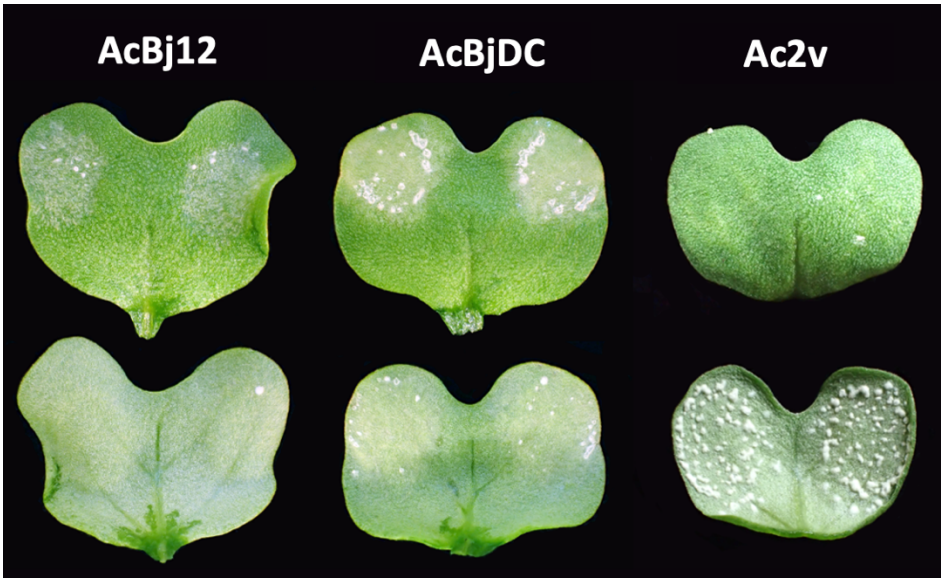


Figure 2.2. Interaction phenotype classes observed on the upper (top row) and lower (bottom row) cotyledon surfaces of *Brassica juncea* accession Donskaja, for UK race 2 isolates AcBj12 and AcBjDC, as well as Canadian race 2 isolate Ac2v. Each presents a distinct level of sporulation ten dpi, with AcBj12 considered as minor virulence, AcBjDC as moderate and Ac2v as major.

2.3.2 Confirming presence of the *BjuWRR1* resistance allele in Donskaja seed stocks. To confirm that the seed stock of Donskaja used as a resistant control in this study contains the *BjuWRR1* resistance allele previously reported by researchers in India (Arora *et al*., 2019), attempts were made to amplify the complete gene from genomic DNA using published gene specific flanking primers. Non-specific amplification was observed, therefore a forward primer designed from just inside the N terminus of exon 1 (DonF1- 5-ATGGCTGACGGAGTTGTGTCG-3) was combined with the original *BjuWRR1* reverse primer (5-GCTCTCACCTTAAATGTTAAAATCGG-3) and a clean product of approximately the correct 5.4 kb length was produced. Sequencing with internal primers captured the complete first and third largest exons, totalling 94.6% of the 2739 bp CDS. Only 147 bp of exon 2 could not be acquired. Sequence was a 100% match to the published *BjuWRR1* allele, containing several SNP and INDEL markers unique to Donskaja when compared to other *B. juncea* accessions. An example of one such unique identifier is illustrated in Figure 2.3.

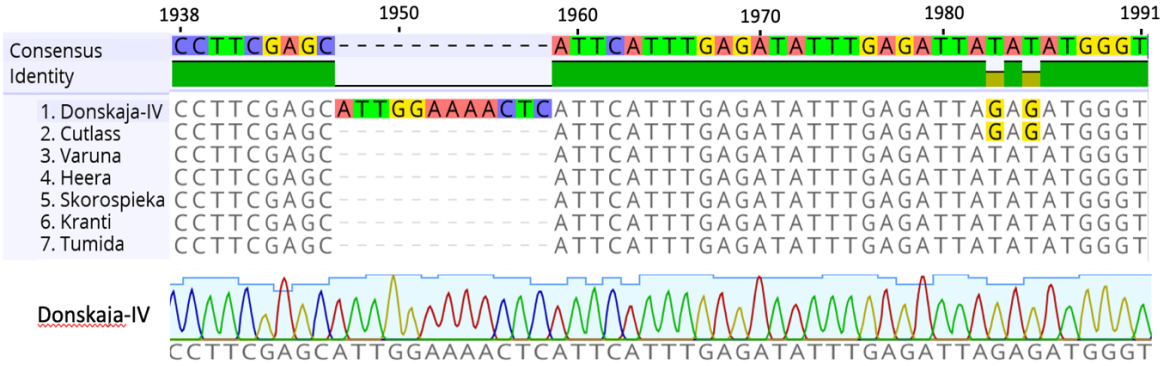


Figure 2.3. Sequence alignment of part of the *BjuWRR1* allele (third exon) for seven *B. juncea* accessions (obtained from the (Arora *et al*., 2019) study), with sanger sequence amplified from the Wellesbourne Donskaja seed stock. The 12 bp INDEL shown here, as well as several other markers captured confirm presence of the *BjuWRR1* resistance allele. Race 2 isolates AcBjDC and Ac2v, which were pathotyped as virulent on this material are therefore able to break the gene-for-gene resistance provided by this well documented *R*-gene.

2.3.3 Identification of genebank *Brassica rapa* accessions segregating for resistance to UK *Albugo candida* race 2 isolates.

Thirteen accessions all exhibited phenotypic variation for resistance and susceptibility to both AcBj12 and AcBjDC (Figures 2.4) and were all thus confirmed as heterogeneous accessions suitable for mapping of WRR loci. Ten of these were advanced for GBS-mapping (Objective 1) by collecting cotyledon samples from a subset of resistant individuals per line, and a bulked sample of cotyledons from at least ten susceptible individuals (Table 2.1). Eight of these produced sufficient datasets required for GBS-mapping of resistance. These accessions represent a range of crop types and one wild species (EH_25). The results from the GBS-mapping experiment subsequently informed design of the following Objective 2 experiment, where five accessions were selected for WGS-BSA mapping.

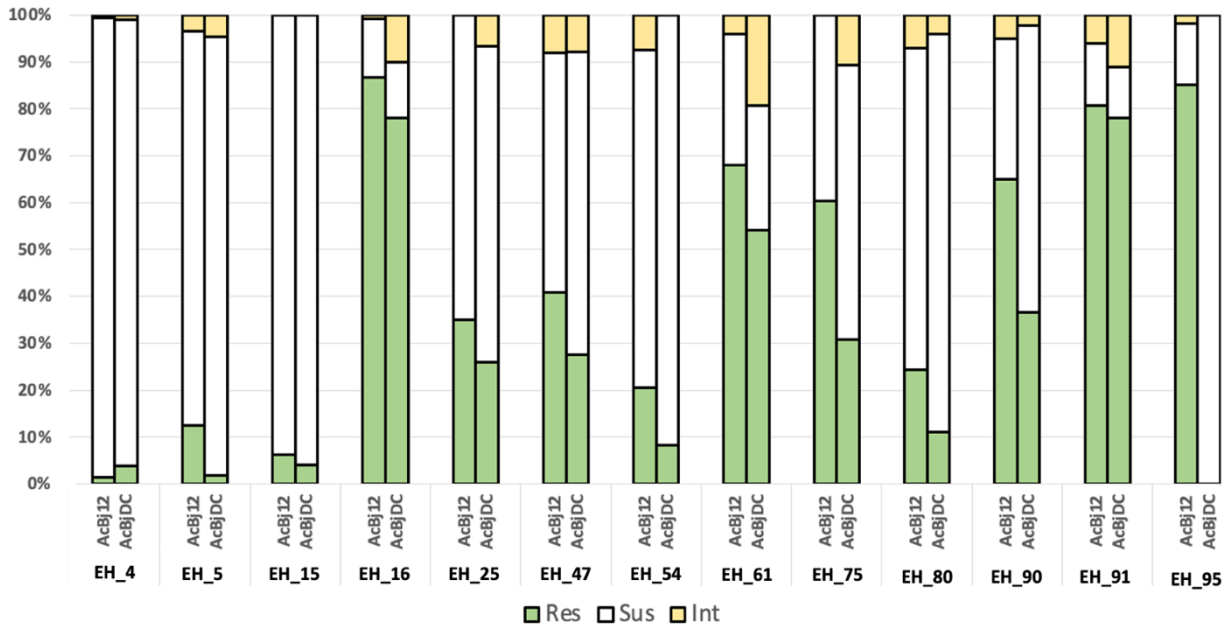


Figure 2.4. Stacked bar plot showing the % degree of phenotypic variation found in each *Brassica rapa* accession when screened with race 2 isolates AcBj12 and AcBjDC respectively. Proportions are an average count of Resistant = no sporulation, Susceptible = full (non-restricted) sporulation within 7-10 dpi and Intermediate = partial and delayed sporulation (10 – 14 dpi) individuals, grown across two propagators per accession. Susceptible controls of Burgonde, Sutton and Cutlass were used. It is observable for each line (except EH_4) that there is a higher degree of susceptibility when tested with AcBjDC compared to AcBj12, suggesting it is the more virulent of the two isolates.

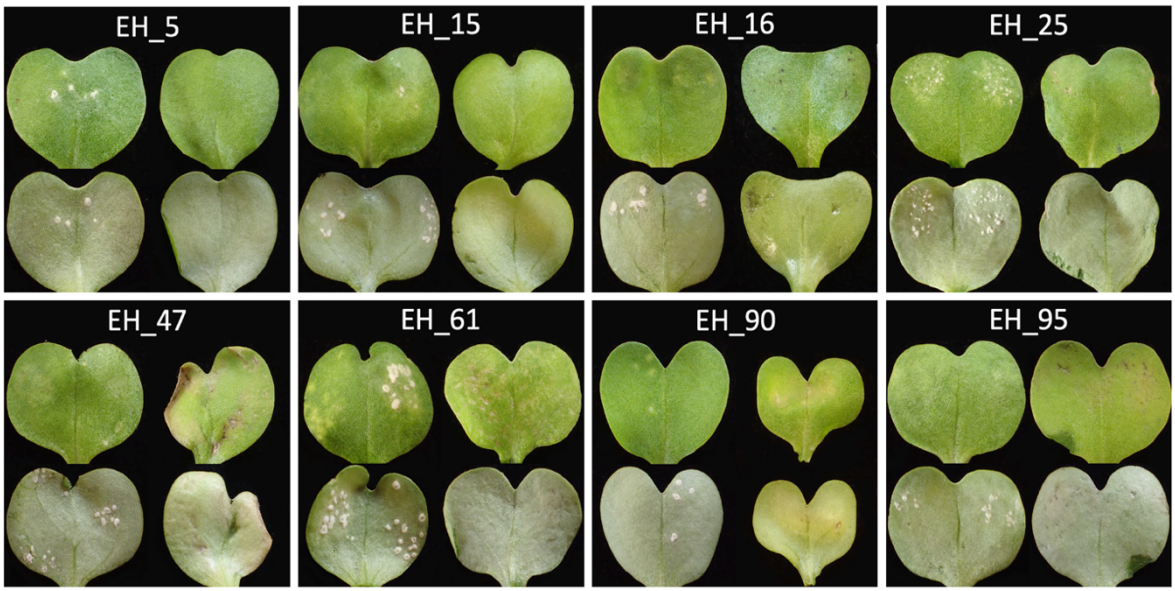


Figure 2.5. Distinct phenotype classes observed on the upper (top row) and lower (bottom row) cotyledon surfaces of the eight *Brassica rapa* accessions selected for GBS-mapping of WRR. Susceptible variants (left) show some degree of sporulation, whilst resistant variants (right) typically presented either some necrosis, or as asymptomatic. Photos were taken ten days post inoculation with the race 2 isolate AcBj12.

Figure 2.6. Adult plant morphology of the eight *Brassica rapa* lines used for GBS-mapping of WRR in this study, demonstrating the wide range of phenotypically variable crop types available. Accessions originate from between Eastern Europe and East Asia.

Table 2.1. List of *Brassica rapa* accessions screened from the UK vegetable genebank that were found to show phenotypic variation (within a seed sample) for resistance/susceptibility following inoculation with two UK isolates of *Albugo candida* race 2 (AcBj12 and AcBjDC). Accessions originate from a wide geographic distribution and represent a range of crop types including one non-domesticated wild type (EH_25). Ten accessions were selected for inclusion for GBS-mapping (Genotyping-by-Sequencing) with a variable number of resistant individuals included. Five of these were subsequently chosen for WGS-BSA (whole genome sequencing for bulked segregant analysis) mapping of major effect white rust resistance loci.

EH no.	HRIGRU ID no.	Subspecies (crop type)	Country origin	GBS	No. res individuals	WGS	No. individuals in susc. Bulk
4	4722	*B. rapa broccoletto* (brocco cima di rapa)	Italy	N	-	-	-
5	5274	*B. rapa broccoletto* (brocco cima di rapa)	Italy	Y	20	Y	37
15	7692	*B. rapa broccoletto* (brocco cima di rapa)	Italy	Y	42	N	27
16	8170	*B. rapa rapa* (turnip)	Syria	Y	10	N	24
25	6699	*B. rapa* sylvestri (wild, non-domesticated)	Algeria	Y	49	Y	35
47	6202	*B. rapa pekinensis* (Chinese cabbage)	Thailand	Y	21	Y	42
54	4714	*B. rapa broccoletto* (brocco cima di rapa)	Italy	N	-	-	-
61	7576	*B. rapa parachinensis* (Choy sum)	China	Y	17	Y	29
75	3116	*B. rapa rapa* (turnip)	Bhutan	N	-	-	-
80	4734	*B. rapa broccoletto* (brocco cima di rapa)	Italy	N	-	-	-
90	4682	*B. rapa perviridis* (Japanese greens)	Japan	Y	21	N	-
91	2488	*B. rapa chinensis* (Chinese cabbage)	Malaysia	N	-	-	-
95	7573	*B. rapa chinensis* (Pak choi)	China	Y	15	Y	25

2.4 Objective 1: mapping WRR loci using GBS data

2.4.1 Identification of resistance-associated SNPs (raSNPs). Approximately 3 million reads were generated from Illumina sequencing of each of the 207 pooled barcoded *B. rapa* samples (eight accessions). The average alignment rate of reads to the *B. rapa* IVFCAASv3 reference genome was 86.53%, with an average single alignment rate of 40.92%, and an average multiple alignment rate of 42.61%. Variant discovery (performed by LGC on all 207 individual sample alignments) produced a single genotype dataframe (minimum allele frequency across all samples > 10%) containing 125680 variants in total. Initial filtering based on the GBS-specific rule set (described in Section 2.26) resulted in the total marker counts for each line shown in Table 2.2. These were then parsed through a succession of filtering steps to reduce the data set and only show markers with a highly conserved association to phenotype, where the resistant genotypes are consistently different to the susceptible bulk genotype. This involved: 1) removal of markers mapped to scaffolds; 2) removal of markers where either of the two susceptible bulk genotypes were missing; 3) removal of markers where the two susceptible bulks (biological replicates) showed differing genotypes; and 4) retaining markers where the proportion of resistant individual genotypes that matched the susceptible genotype (MatchCount) was either ≤ 10% (non-conservative filter) or ≤ 5% (conservative filter).

Primarily, this approach was undertaken to identify dominant raSNPs with a homozygous susceptible bulk genotype vs an alternative heterozygous or homozygous resistant genotype. Non-conservative filtering resulted in a total of 117 raSNPs across all lines, whilst a conservative raSNP count identified a total of 36 markers. An attempt was made to also identify recessive raSNPs (with heterozygous susceptible bulks and homozygous resistant individuals). However, even with the maximum possible MatchCount stringency (parameter = 0) only lines EH_15 and EH_25 produced < 100 raSNPs each (Table 2.2). These two lines were represented with the highest number of resistant individuals (42 and 49 respectively). The relationship between the number of raSNPs identified per line and the number of resistant individuals included for GBS-mapping can be seen in Figure 2.7. To estimate an average error rate in genotype calling, disagreement between the total calls of the two biological replicates (two susceptible bulks) was calculated at 13.6%.

Table 2.2. Summary of GBS marker filtering, for the identification of resistance-associated SNPs (raSNPs). Filtering steps previously described were applied to identify both dominant (non-conservative = MatchCount ≤ 10%; conservative = MatchCount ≤ 5%;) and recessive raSNPs (MatchCount = 0%)) from a database containing GBS-mapping markers for eight *Brassica rapa* genebank accessions.

Line	EH_5	EH_15	EH_16	EH_25	EH_47	EH_61	EH_90	EH_95	Average	Average% decrease
Number Res. individuals	20	42	10	49	21	17	21	15		
Total markers	86872	75206	99423	77489	87602	91347	86947	76602	**85186**	
Scaffolds	82128	71318	94103	73494	82648	86108	81993	72495	**80536**	5.5
Sus_missing	77989	68462	89254	69893	80825	84612	80262	73805	**78138**	8.3
Sus_similarity	66982	57038	77477	60837	69664	75770	71023	61462	**67532**	20.7
Rec MatchCount = 0%	346	25	1113	30	103	325	202	169	**287**	99.663
Dom MatchCount ≤ 10%	27	13	37	3	10	13	10	4	**15**	99.983
Dom MatchCount ≤ 5%	6	1	6	3	4	4	8	4	**5**	99.995

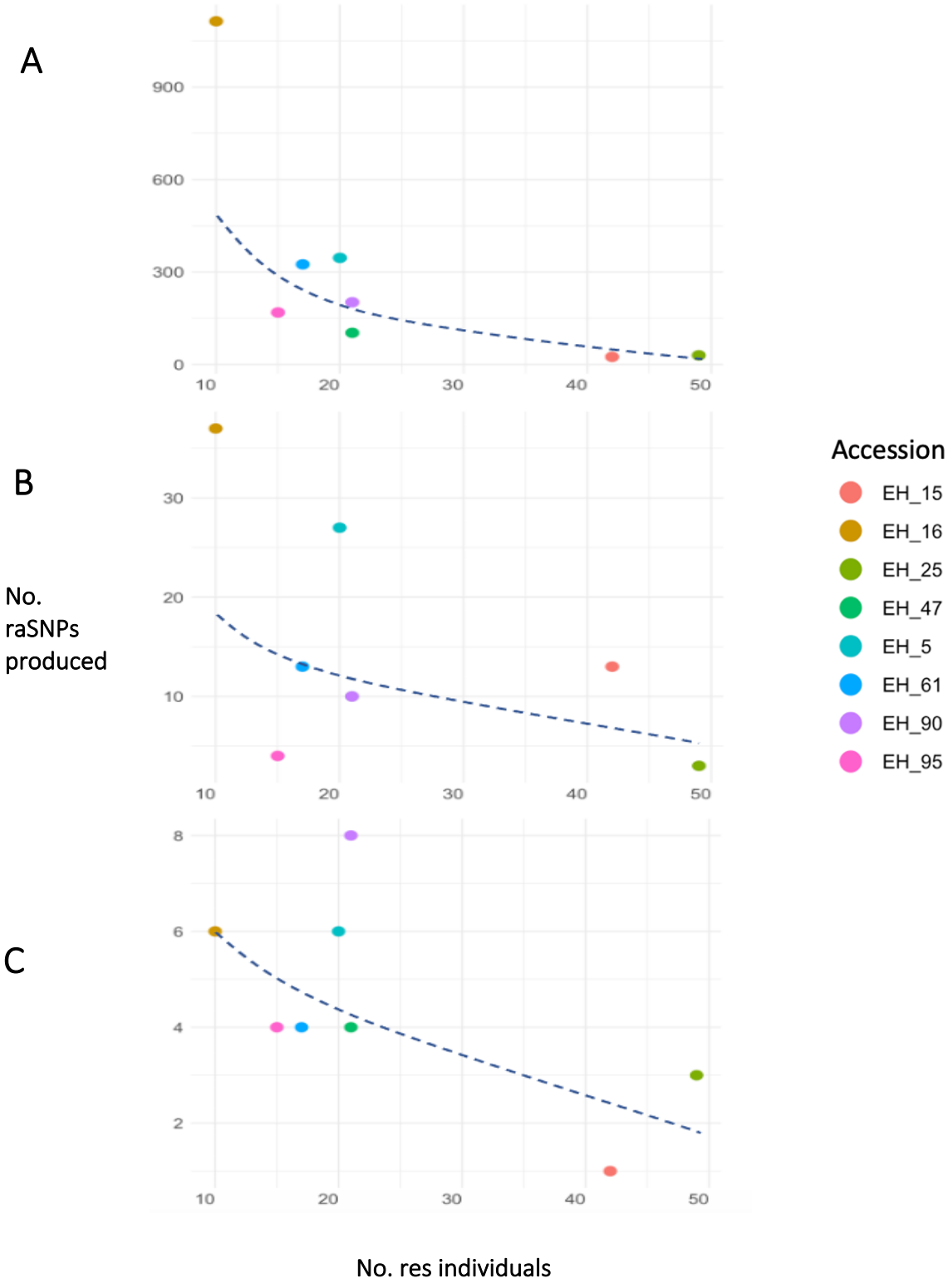
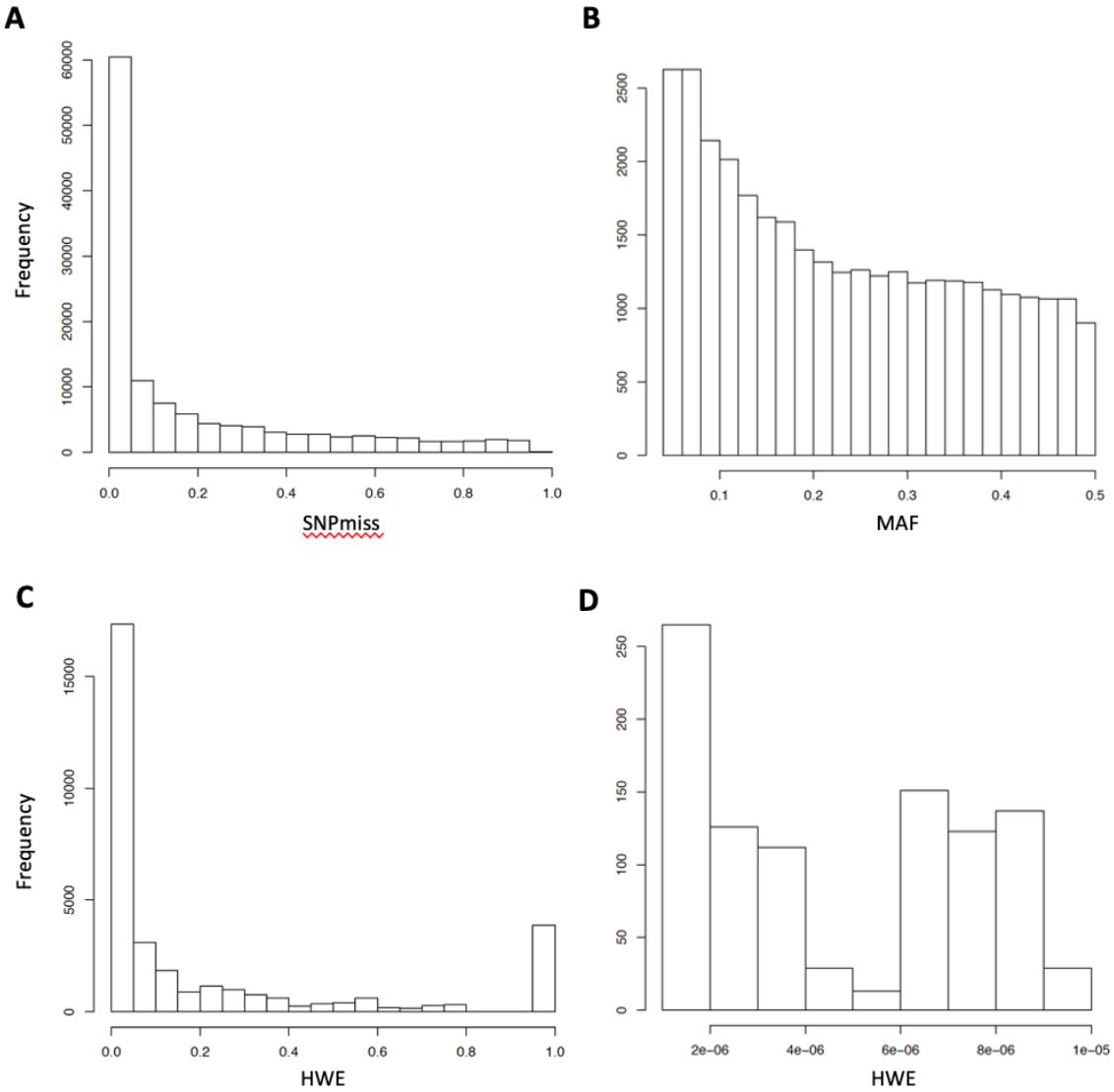


Figure 2.7. Response between the number of resistant individuals provided per line for GBS-mapping and the number of outputted resistance-associated SNPs (raSNPs) including: **A)** Recessive markers filtered for MatchCount = 0, with each point representing one of the eight *Brassica rapa* accessions. ; **B)** non-conservative dominant markers (MatchCount ≤ 10%); and **C)** Conservative dominant markers (MatchCount ≤ 5%). Logarithmic lines of best fit indicate a negative correlation between these two factors.

2.4.2 Characterisation of population structure, linkage disequilibrium and phylogenetic relatedness amongst *Brassica rapa* accessions using GBS data

Prior to GBS-mapping, a subset of GBS SNP data (containing five randomly selected individuals per accession) was assembled into a GBS-subset dataframe to make an assessment of population structure both within and between the geographically diverse *B. rapa* populations. In addition to the eight accessions used for GBS-mapping, data was also available for accessions EH_75 (Bhutan) and EH_91 (Malaysia) which have been added to the genome analyses conducted here. A GWAS QC pipeline developed by (Marees *et al*., 2018) was applied to subset for high quality markers, which were then inputted into the model-based program fastSTRUCTURE (Raj *et al*., 2014). SNPs were filtered based on the following factors: **1)** Delete SNPs with high levels of missing data (SNP missingness > 0.02); **2)** Remove SNPs with a low minor allele frequency (MAF) (MAF < 0.05); and **3)** Delete SNPs which are not in Hardy-Weinberg equilibrium (HWE) (HWE < 1e-6). Parameter setting were determined by visualising histogram frequencies for calculations made at each step (Figure 2.8). Filtering of the GBS-subset (containing 124000 SNPs) produced a final set of 31883 SNPs distributed across the *B. rapa* genome.

Population structure was estimated with the 31883 QC filtered SNPs using FastSTRUCTURE (Raj *et al*., 2014), which uses a variational Bayesian framework for posterior inference to rapidly infer population structure from large SNP genotype data. The algorithm was run independently for population sizes (K) of 1:10 to obtain a reasonable range of values for the appropriate model complexity required to explain structure in the data. The model complexity that maximised marginal likelihood) was 7, and the model components (populations) used to explain structure in the data was K = 6, which ran for 30 iterations. The degree of admixture inferred by FastSTRUCTURE can be visualised in the distruct plot (Figure 2.9), generated from the mean of the variational posterior distribution over admixture proportions. Results show clear differences in structure between accessions relative to individuals from the same accession. Accessions with enough similarity to be included in the same population include an east European pairing (EH_15 from Italy and EH_25 from Algeria), and an east Asian pairing (EH_47 from Thailand and EH_61 from China). Interestingly, the Japanese accession EH_90 displayed significant admixture, with genetic contributions from three Asian populations and an east European population (represented by EH_5 and EH_25).



.

Figure 2.8. Histograms of SNP frequencies from the GBS-subset (five individuals per accession) dataframe, QC filtered for input into FastSTRUCTURE. SNPs were removed with respect to: A) SNP missingness > 0.02; B) MAF < 0.05; C) HWE < 1e-6. Graph D provides a zoomed in visual of SNPs strongly deviating from HWE, which were subsequently removed from the dataset.

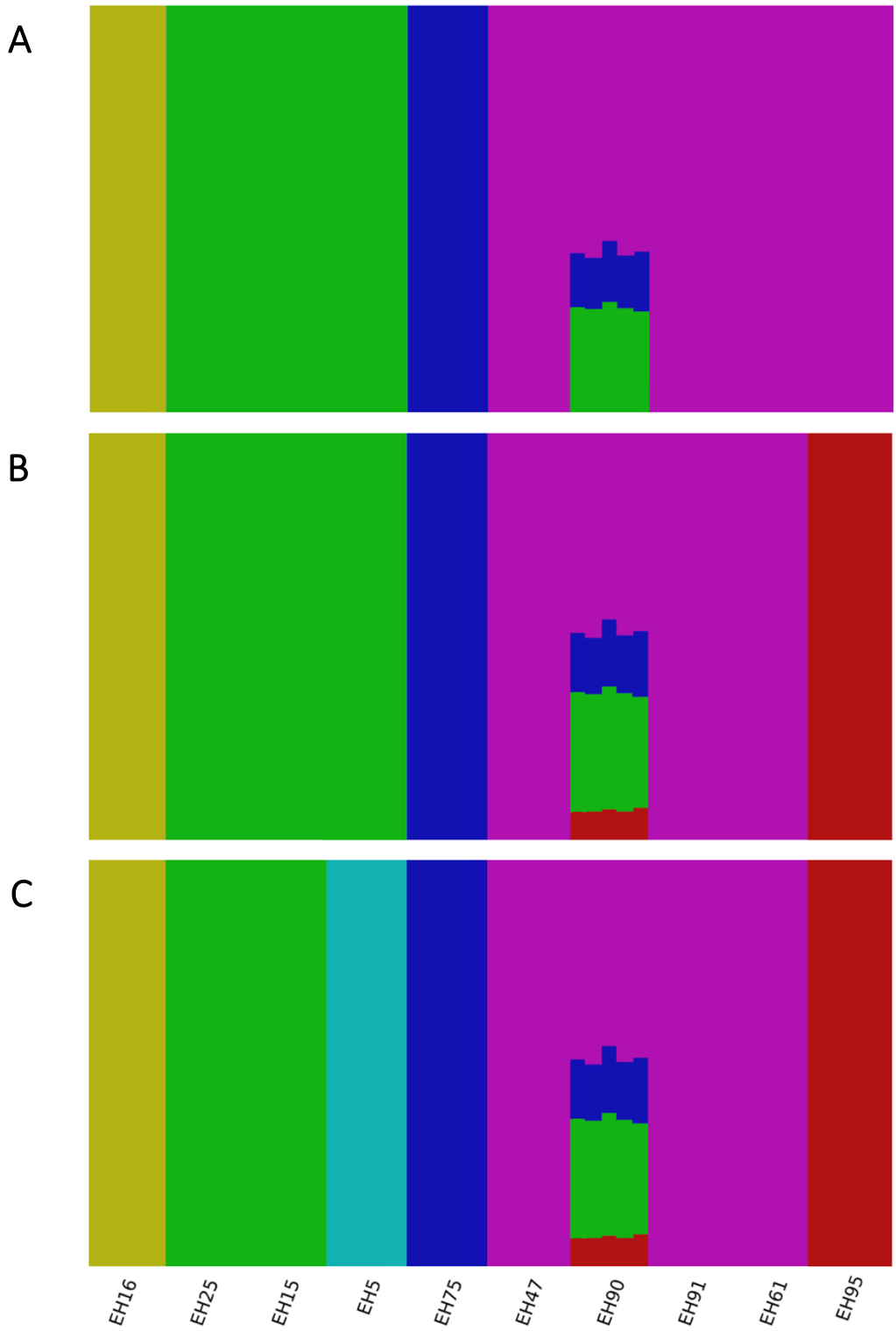


Figure 2.9. Distruct plots showing relative population structure of GBS data from ten *B. rapa* accessions (five individuals included per accession). Analysis was performed using fastSTRUCTURE, based on an optimal population number of K = 6 (C), though K = 4 (A) and K = 5 (B) have also been provided for comparison. Each individual is represented by a vertical bar, where colour indicates the estimated membership fractions for K = *n* clusters (or sub populations). The admixture seen in accession EH_90 suggesting historical breeding between at least four different structural classes.

Patterns of linkage disequilibrium (LD) were estimated from the same GBS-subset used for structure analysis, with the LD analysis function in Tassel 5 (Bradbury *et al*., 2007). A sliding window of 50 SNPs was applied making 1592875 comparisons across the dataframe to calculate r^2 statistics. The r^2 value is the square of the correlation coefficient between two indicator variables which represent the presence absence state of alleles as two different loci (Hill & Robertson, 1968). The resulting LD plot shown in Figure 2.10 highlights a 300 kb window on chromosome 6, showing haplotype blocks of markers in LD as determined by the r2 statistic. To visualise LD decay over distance, a custom R script (https://www.biostars.org/p/300381/#300423) using plink (v1.90, (Purcell *et al*., 2007)) was adapted to measure and plot the mean r2 within 10 kb intervals (Figure 2.11).

A single resistant individual from each of the ten accessions was selected to assess phylogenetic relatedness. Analysis was undertaken using Tassel 5 (Bradbury *et al*., 2007) and a neighbour joining clustering method to produce a distance matrix and the phylogenetic tree seen in Figure 2.12. A total of 31899 GBS markers across all ten *B. rapa* chromosomes was used to compile the tree. Two divergent clades were produced that distinguish the Asian accessions (clade 1), from the Eastern European accessions (clade 2).

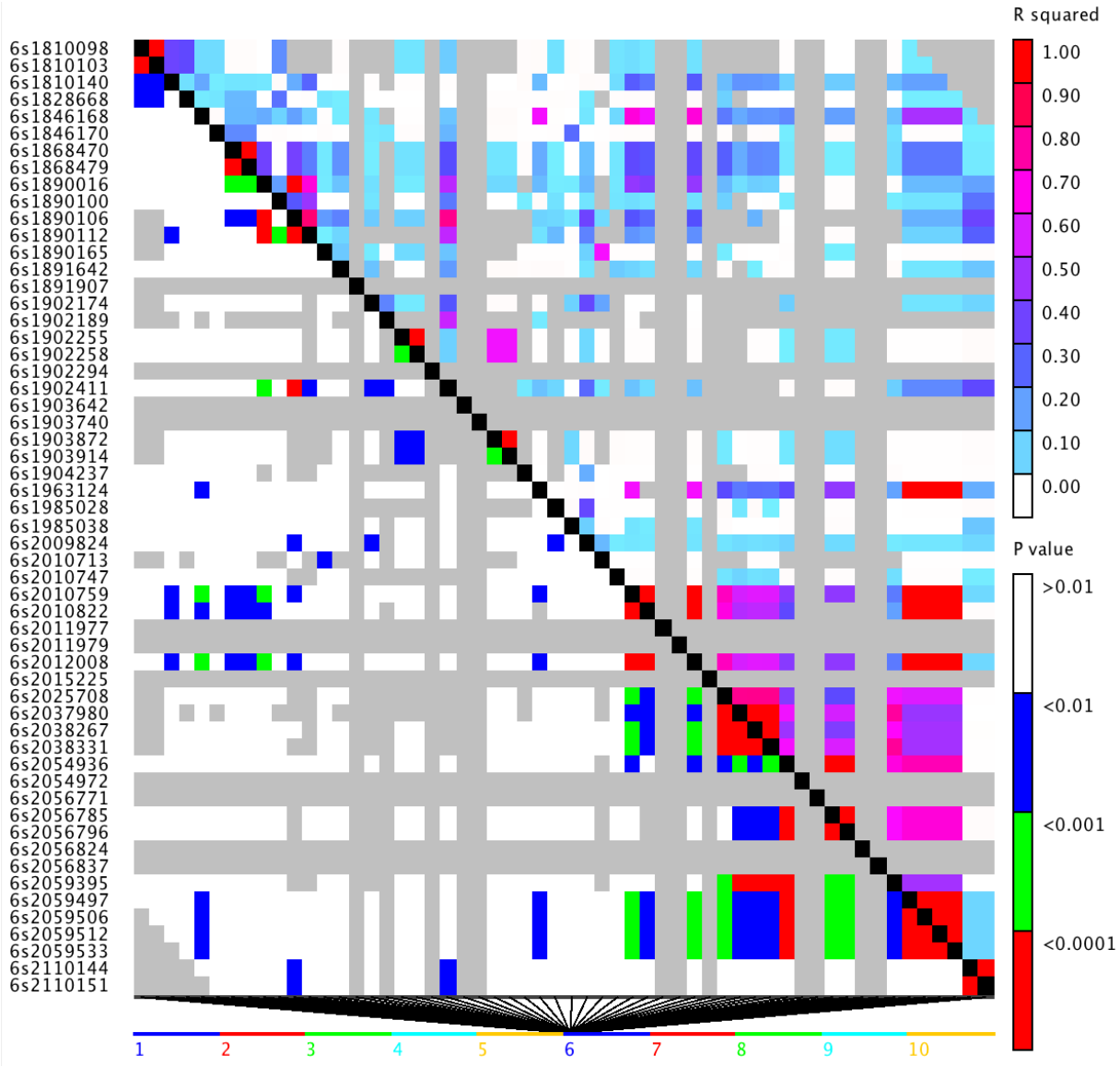


Figure 2.10. Visualisation of LD in *Brassica rapa* accessions (cumulative GBS dataset) for a 200 kb window on chromosome 6. The colour spectrums above and below the black diagonal line represent the r^2 (above) and the *p*-value (below) scores respectively. Haplotype blocks can be seen in red, which indicate regions where there has been little historical recombination.

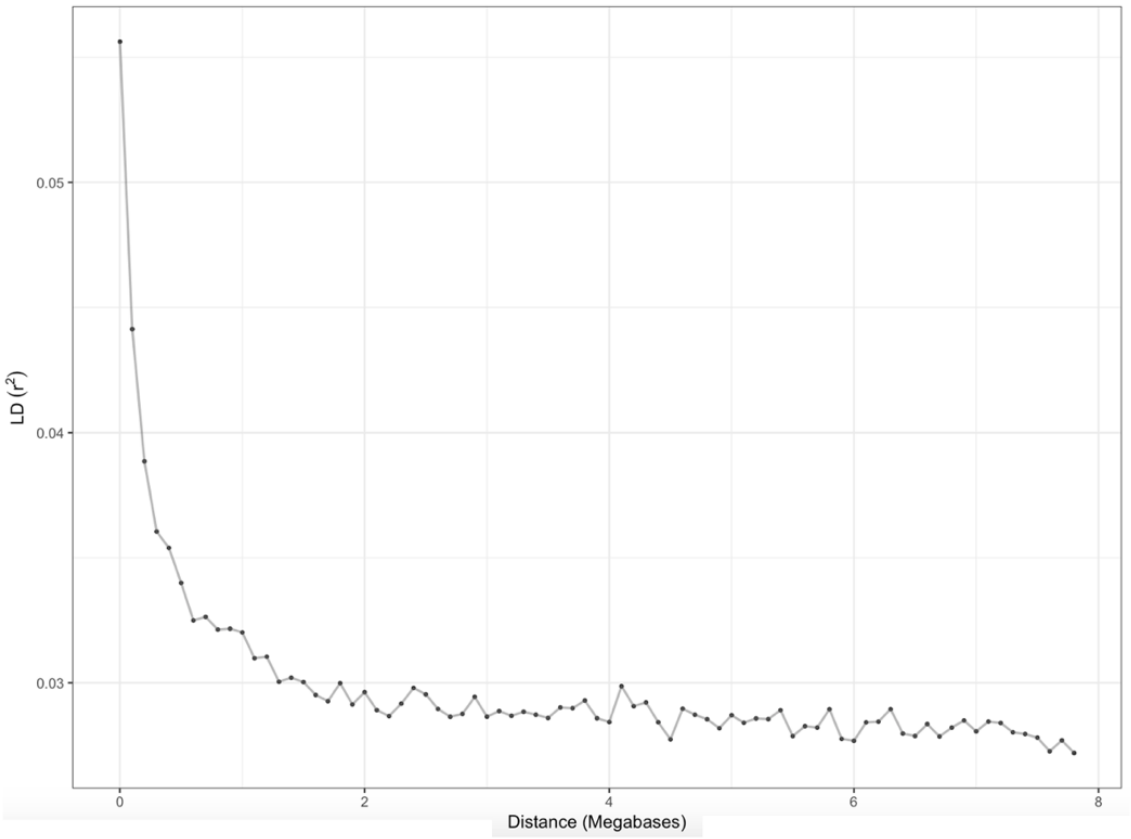


Figure 2.11. Genome-wide decay of the linkage disequilibrium (LD) in *Brassica rapa* accessions (cumulative GBS dataset).

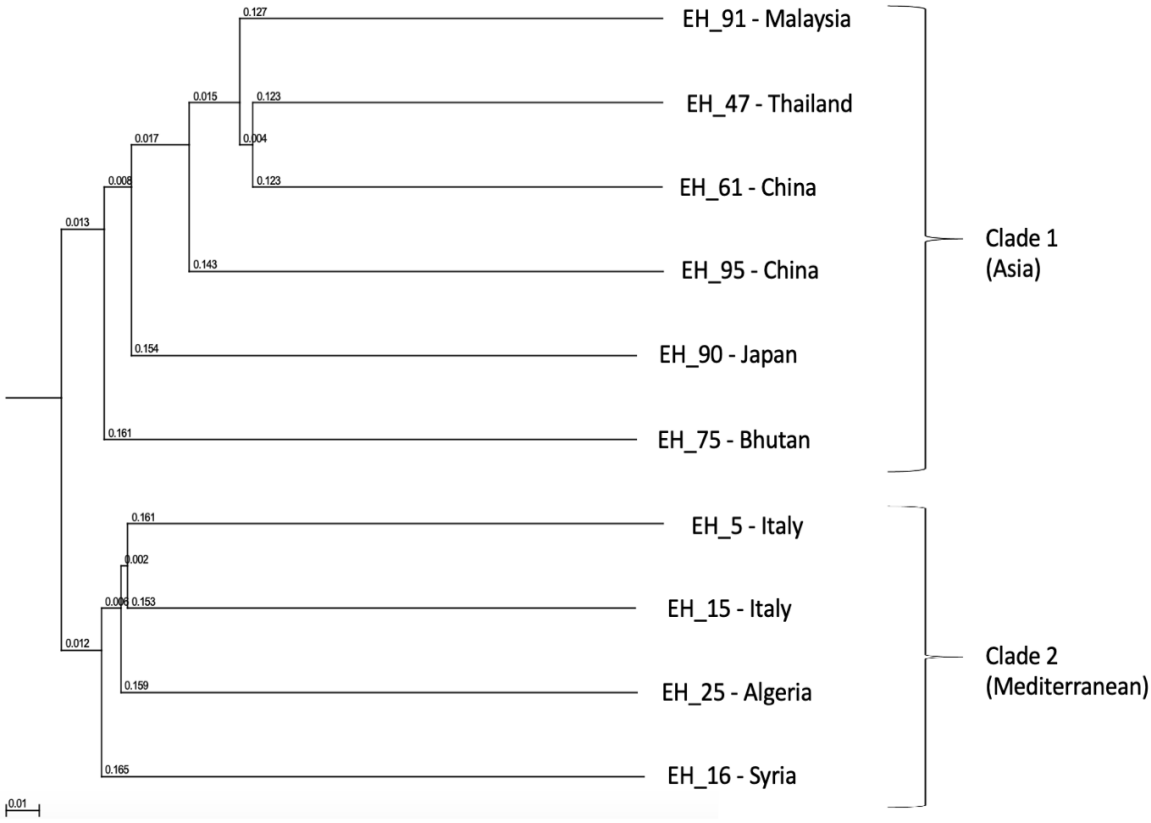


Figure 2.12. Neighbour joining tree constructed from GBS sequence of the ten *Brassica rapa* genebank accessions. SNP data for one individual per accession was used to collate 31899 markers in total, spanning all ten chromosomes. The scale bar indicates 0.01 substitutions per site. Accessions can be seen to cluster into two main clades which reflect their geographic origins of domestication, ranging longitudinally from between Algeria (EH_25) and Malaysia (EH_91).

2.4.3 Identifying *R*-genes relative to positions of GBS-mapping raSNPs. DRAGO 2 software was able to detect possible *R*-genes genes from any combination of the following resistance associated domains; coiled-coil (C), kinase (K), leucine rich repeat (L), nucleotide binding domain (N), protein (P) and Toll-interleukin region (T). A total of 1002 *R*-genes were detected across the reference assembly that exclusively contained these domains (Figures 2.13 and 2.14). Approximately 50% make up transmembrane receptors including receptor like kinases (RLK) and proteins (RLP), with the remaining predominantly of the intracellular receptor class which includes toll-interleukin region or coiled-coil NLR (TNL and CNL) genes.

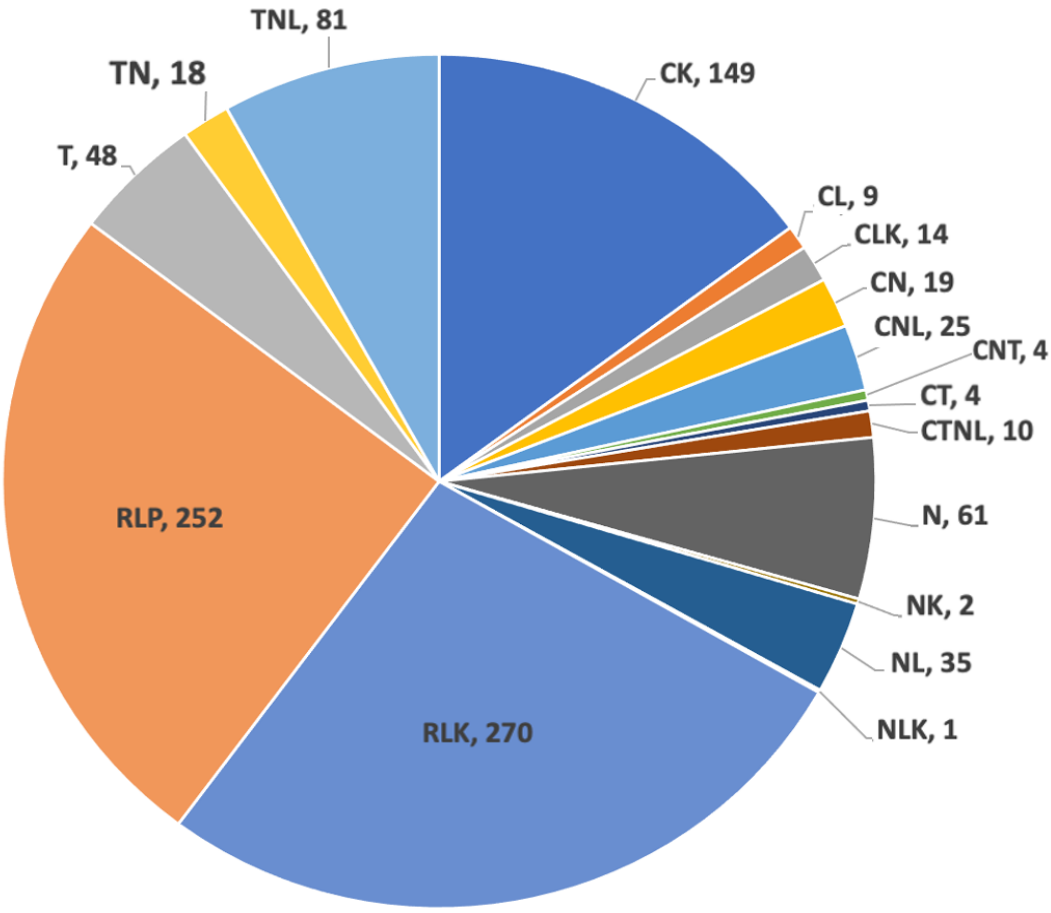


Figure 2.13. Resistance genes (*R*-genes) annotated in the *Brassica rapa* IVFCAASv3 reference assembly using DRAGO 2 hidden Markov model software. In total, 1002 *R*-genes were identified, with approximately 50% making up transmembrane receptors (receptor like kinases (RLKs) and proteins (RLPs)). The remaining 50% are predominantly of the intracellular receptor class, with the number of major class TNL and CNL receptors totalling 106. Genes make up combinations of domain classes that include: coiled-coil (C), kinase (K), leucine rich repeat (L), nucleotide binding domain (N), protein (P) and Toll-interleukin region (T).

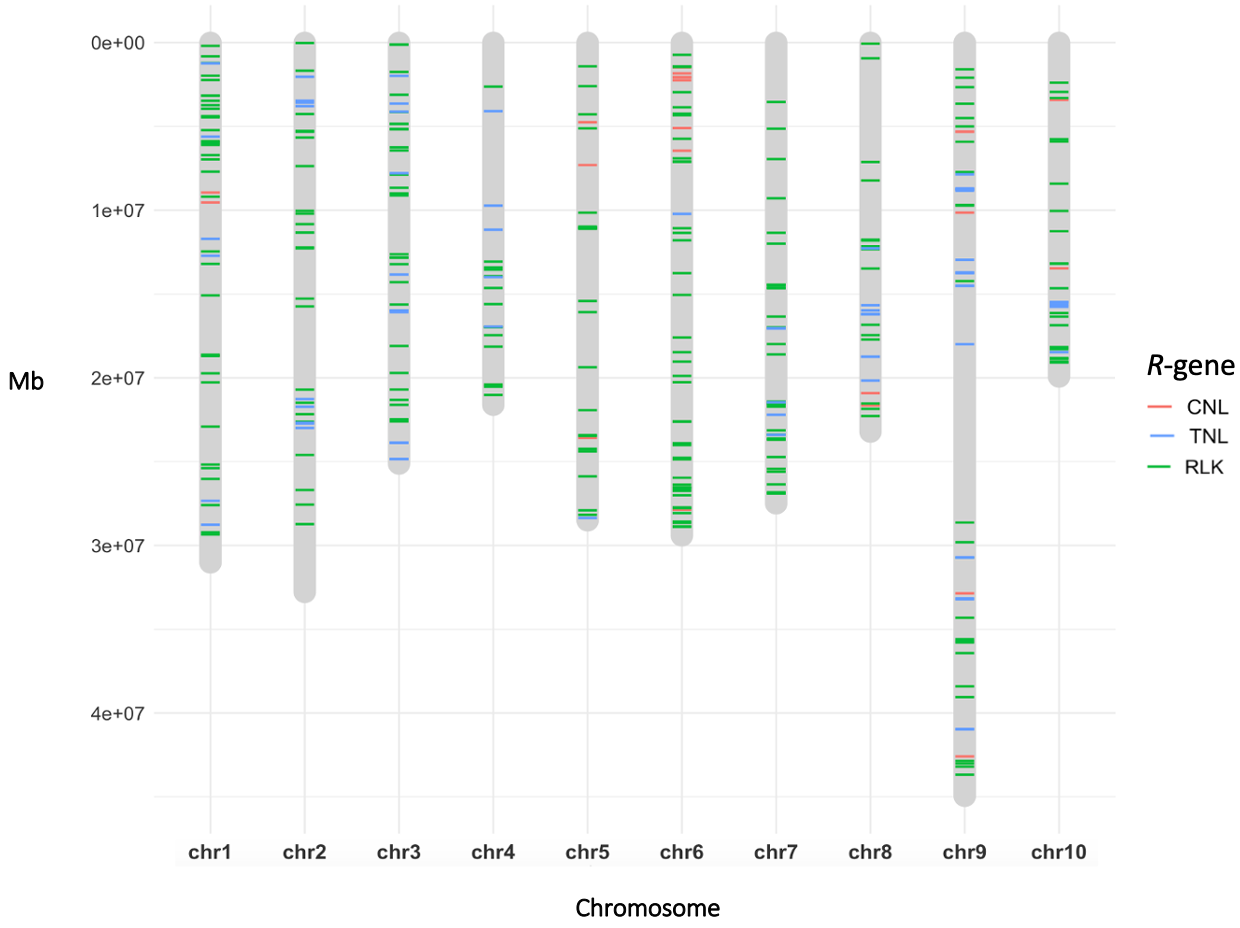


Figure 2.14. Physical positions of annotated *R*-genes across the *Brasscia rapa* IVFCAASv3 reference genome. These represent the major classes that encode CC-NB-LRR (CNL), TIR-NB-LRRs (TNL) and receptor-like kinases (RLK).

2.4.4 Physical mapping of GBS-mapped raSNPs against the *Brassica rapa* reference.

A custom R script (Appendix Script 1) was used to physically position filtered GBS-mapped raSNPs from eight *B. rapa* accessions (Table 2.2) on the ten chromosomes of the *B. rapa* IVFCAASv3 reference assembly. Collectively, all chromosomes except nine and ten contained mapped raSNPs.

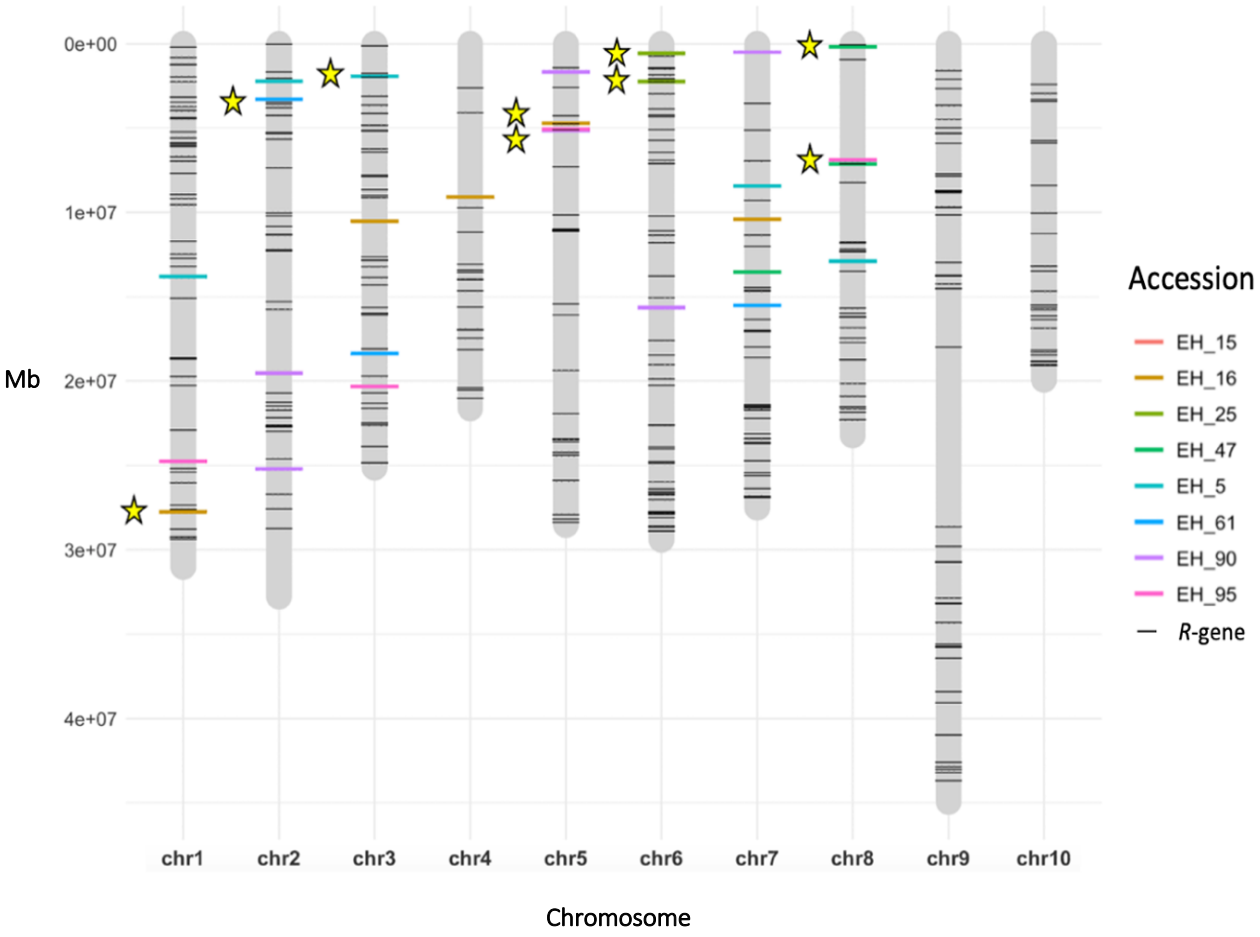


Figure 2.15. Distribution of the 37 conserved (MatchCount ≤ 5%) dominant raSNPs (AcBj12) across the *Brassica rapa* IVFCAASv3 reference genome, for each accession, and their positions relative to *R*-genes (CNL, TNL and RLKs). Yellow stars denote the 14 raSNPs that are within a 200 kb physical distance of an *R*-gene.

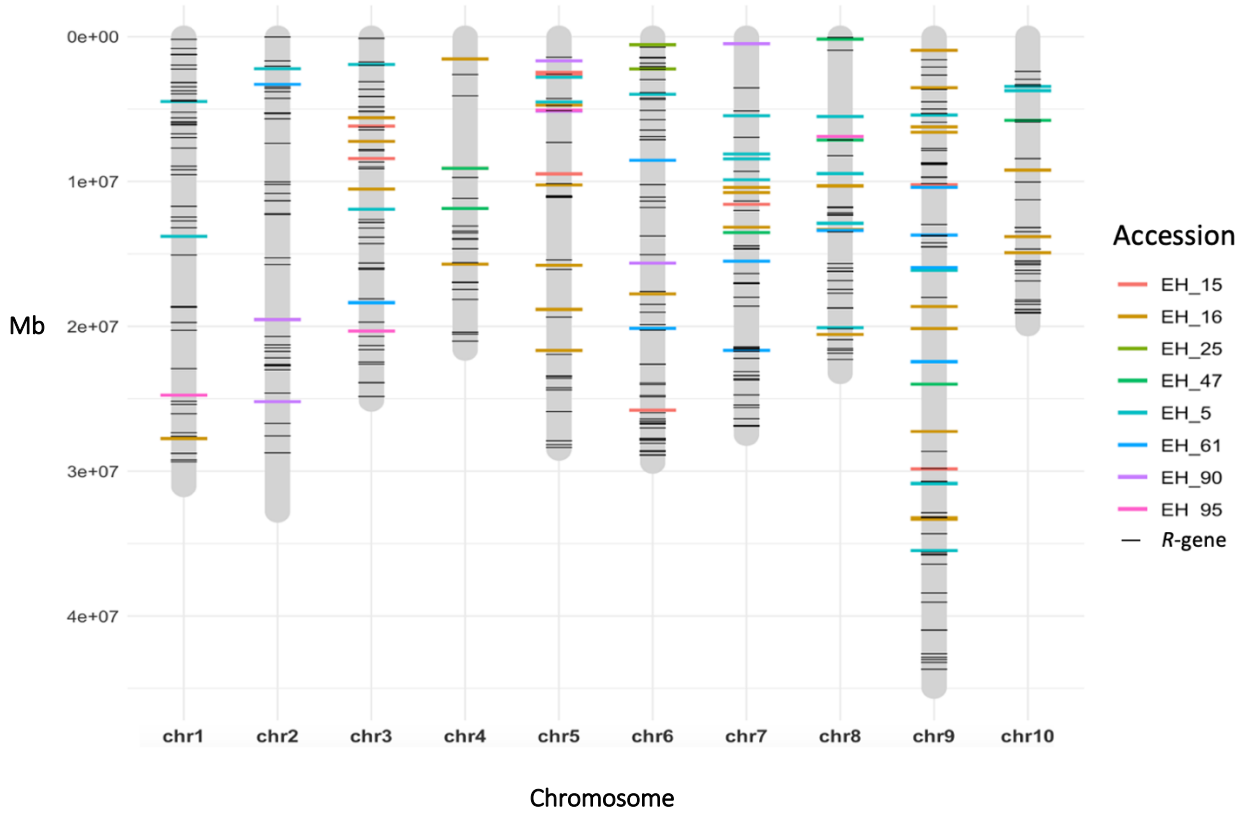


Figure 2.16. Distribution of 117 non-conserved (MatchCount ≤ 10%) dominant raSNPs (AcBj12) across the *Brassica rapa* IVFCAASv3 reference genome for eight genebank accessions. Positions can be seen relative to major *R*-gene classes (CNL, TNL and RLKs).

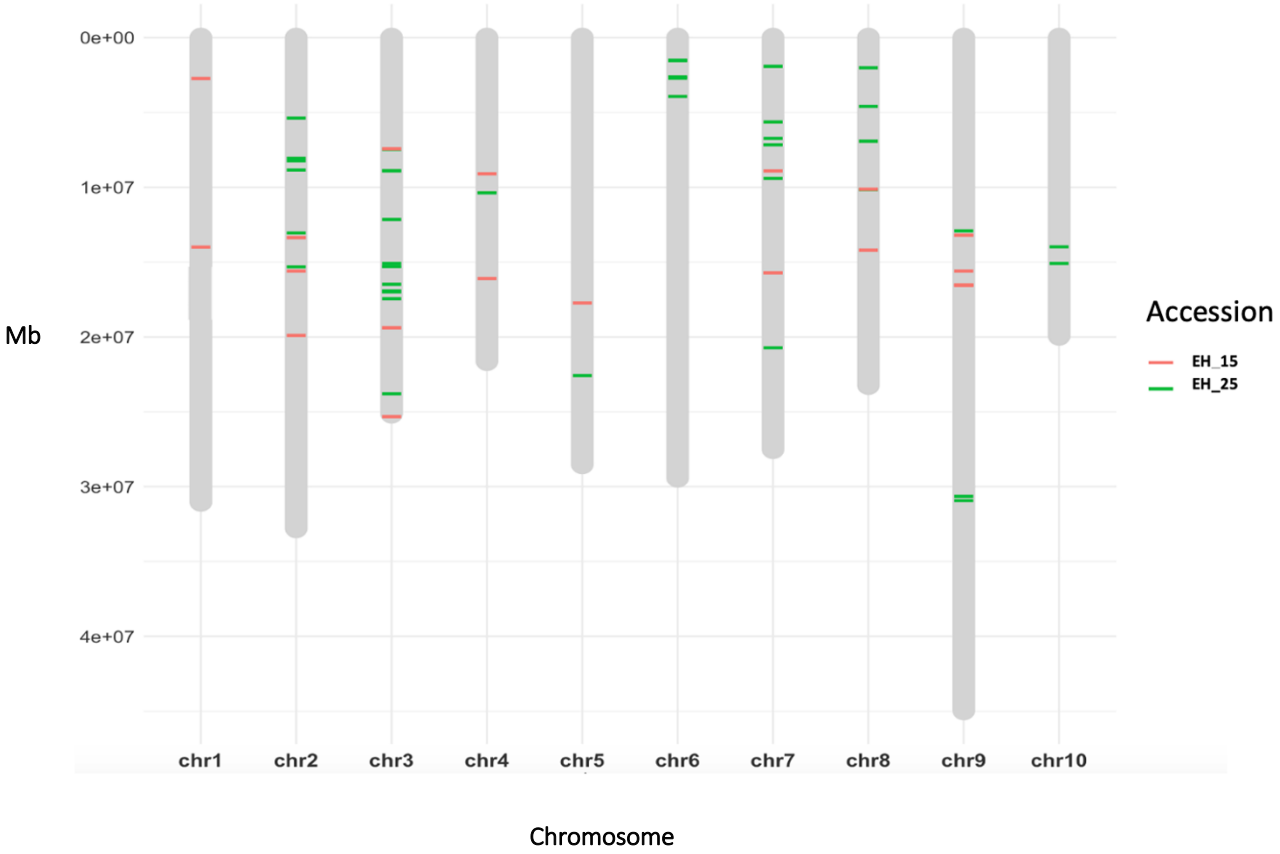


Figure 2.17. Distribution of recessive raSNPs for lines EH_15 and EH_25 across the *B. rapa* IVFCAASv3 reference assembly, filtered for maximum association to phenotype (MatchCount = 0). The remaining lines, which were each represented by ≤ 21 resistant individuals, generated high numbers (> 100) of raSNPs, and so have not been represented here.

Table 2.3. List of conserved dominant GBS-mapping raSNPs that are < 200 kb from a predicted *R*-gene (class TNL, CNL and RLK).

Chr	Line	raSNP pos	*R*-gene < 200 kb	Class
A01	EH_16	27,737,747 27,750,382	BraA01g041090	RLK
A02	EH_5	2,215,025	BraA02g004210	TNL
A03	EH_5	1,922,446	BraA03g004100 BraA03g004600	RLK TNL
A05	EH_16	4,726,668	BraA05g008990	CNL
	EH_95	5,074,313	BraA05g009590	RLK
	EH_90	5,145,355 5,145,405	BraA05g009590	RLK
A06	EH_25	544,426 574,781	BraA06g001180	RLK
		2,230,620	BraA06g003420 BraA06g003430 BraA06g003770	CNL CNL CNL
A08	EH_47	168249 168276	BraA08g000110	RLK
		7130746	BraA08g008060	RLK

2.5 Objective 2: mapping WRR loci using WGS-BSA data

2.5.1 Referenced-based SNP identification from WGS of *B. rapa* tissue bulks. Whole genome sequences (WGS) of pooled DNAs were generated using paired end (150 bp) sequencing on the Illumina Novaseq6000 platform. Sequencing generated a total of 466.6 Gb of raw data, with an average of 110 and 116 M PE reads for the resistant and susceptible samples respectively. Low quality reads and adapters were removed before checking read quality using FASTQC (Andrews, 2010). High-quality sequences were aligned and mapped to the *B. rapa* IVFCAASv3 reference assembly (Wang *et al*., 2011) downloaded from http://brassicadb.org/brad/ using BWA with default parameters. Across all samples, the average coverage of aligned sequence reads to the reference was 68.2%, with an average of 95.6% breadth of the reference genome covered by mapped reads. The average alignment rate of reads to the reference was 84.18%, with an average of 75.12% mapping in pairs. Mapping statistics for individual lines can be viewed in Table 2.4.

Reference-based SNP identification was applied as described in Section 2.2.6. SNP datasets were then merged per accession for each resistant and susceptible bulk to produce the SNP counts seen in Table 2.5.

2.5.2 WGS based BSA mapping of WRR QTL in *B. rapa* genebank accessions. To map WRR QTL, WGS-BSA was performed using QTLseqr (Mansfeld & Grumet, 2018) and the G' approach (Magwene *et al*., 2011). Consolidated resistant and susceptible SNP tables were passed into QTLseqr using the function importFromGATK, which also calculates the following parameters for each SNP in the dataframe:

$$\textit{Reference allele frequency} = \frac{\textit{Ref allele depth}_{\textit{HighBulk}} + \textit{Ref allele depth}_{\textit{LowBulk}}}{\textit{Total read depth for both bulks}}$$

$$\textit{SNP-index}_{\textit{per bulk}} = \frac{\textit{Alternate allele depth}}{\textit{Total read depth}}$$

$$\Delta\textit{(SNP-index)} = \textit{SNP-index}_{\textit{HighBulk}} - \textit{SNP-index}_{\textit{LowBulk}}$$

Resistant bulks were assigned as the high bulk and susceptible bulks as the low. To select for high-confidence SNPs for analysis, the following six filtering steps were applied to each dataset: 1) reference allele frequency ($0.3 \leq$ REF_FRQ ≤ 0.7); 2) total sample read depth (total DP ≥ 60); 3) total sample read depth (total DP ≤ 400); 4) per sample read depth (DP ≥ 30); 5) GATK genotype quality (GQ ≥ 99); and 6) difference between bulks (≤ 100). The number of SNPs filtered at each step has been reported per accession in Table 2.5. Filtering here serves to remove extremely low and high coverage (from repetitive regions) SNPs, either in both bulks (Steps 1 and 2), and/or in each bulk independently. The largest reduction came from filtering the reference AF (SNPs that are over or under-represented in both bulks), where an average of 69.26% of SNPs were removed from the dataset.

The SNP index and G statistics were calculated for each individual SNP as described by (Magwene *et al*., 2011). Tricube smoothed Δ(SNP-index) and G values (G' value) were calculated within a window size spanning 1.0 Mb of genomic region and were plotted against all ten *B. rapa* chromosomes (Figure 2.18). Here, the G' statistic functions as a weighted moving average across neighbouring SNPs, accounting for linkage disequilibrium (LD), whilst also minimising noise attributed to SNP calling errors (Mansfeld & Grumet, 2018). G' values calculated for each SNP are weighted by physical distance to the focal SNP, decreasing in value as they get closer to the window edge. Significance thresholds (*p*-values), and genome-wide Benjamini-Hochberg false discovery rate (FDR) (Hochberg, 2016) adjusted *p*-values (*q*-values) were estimated from the null distribution of the G', which assumes there is no QTL linked to the SNP. The tricube smoothed G' values from the analysis show significant peaks across the FDR (q) threshold of 0.001 in all accessions. In total, this produced 26 QTLs, though due to noise in the background of the EH_47 output, only the four most prominent G' peaks on A05 and A07 will be considered for this accession. This leaves a total of 16 mapped QTLs (EH_5 = 3, EH_25 (AcBj12) = 1, EH_25 (AcBjDC) = 1, EH_47 (most prominent QTLs) = 4, EH_61 = 3 and EH_95 = 4). Chromosomes A08 and A10 are the only two not to contain identified QTLs for any accession. The genomic regions mapped by QTLs (across all accessions) varied from the smallest of 250 kb (EH_47 - A05) to the largest of 7.22 Mb (EH_95 – A02), with the average interval spanning 2.44 Mb. Statistics inferred from QTLs of each accession have been documented in Table 2.6.

Amongst all 16 QTLs, EH_25_12.A06 and EH_95.A02.2 show the highest G' peaks (11.4 and 11.3 respectively), indicating major effect QTLs for WRR. Interestingly, these genomic regions on both chromosome A02 and A06 have been mapped independently by different accessions. The A02 region (18670297 – 21646283 / 2.96 Mb), termed BraA02 is spanned by both EH_61 (EH_61.A02) and EH_95 QTLs (EH_95.A02.2), whilst the A06 region (948017 – 2715701 / 1.76 Mb) termed BraA06 is spanned three times, twice by EH_25 (AcBj12 and AcBjDC) and once by EH_5. The overlap of mapped QTLs here indicates the possibility of shared resistance loci in these regions. This is supported by the relative geographic origins of the accessions, with the A02 region mapped using Chinese accessions and the A06 region mapped using accessions from Italy (EH_5) and Algeria (EH_25). Mapping the A06 region twice in EH_25 (EH_25_12.A06 + EH_25_DC.A06) using both race 2 *A. candida* isolates (AcBj12 and AcBjDC) indicates a potentially shared resistance locus capable of preventing infection to both isolates. Other regions of interest include where clear G' peaks are intercepted by short non-significant gaps, such as on A02 (EH_95) and A05 (EH_47). These examples suggest multiple QTLs that are linked in coupling phase.

Table 2.4. Paired-end read counts in millions (M) from WGS at 90x sequencing coverage and alignment statistics (against the *B. rapa* IVFCAASv3 reference assembly) for each accession. Statistics are averaged between the resistant and susceptible files and represent average read depth across the genome, average breadth of coverage and the % of reads that mapped to the refer **Table 2.5.** SNPs filtered using the QTLseqr pipeline prior to association mapping of WRR. Accessions were filtered by; 1) Reference allele frequency (REF_FRQ); 2) Total sample read depth (DP); 3) Total sample DP; 4) Per sample DP; 5) Genotype quality (GQ); 6) Difference between bulks (DBB).

	EH_5	EH_25_12	EH_25_DC	EH_47	EH_61	EH_95
Res bulk reads (M)	163	108	114	103	102	103
Sus bulk reads (M)	109	127	105	101	101	102
Depth coverage (x)	63.3	76.4	70.3	66.9	65.4	66.7
Breadth coverage	94.6	96.2	95.9	97.1	96.1	93.9
% mapped to ref.	79.1	79.9	85.4	88.9	87.8	83.8

Table 2.5. SNPs filtered using the QTLseqr pipeline prior to association mapping of WRR. Accessions were filtered by; 1) Reference allele frequency (REF_FRQ); 2) Total sample read depth (DP); 3) Total sample DP; 4) Per sample DP; 5) Genotype quality (GQ); 6) Difference between bulks.

Step	Parameter	EH_5	EH_25_12	EH_25_DC	EH_47	EH_61	EH_95	Mean	Mean %
1	0.3 ≤ REF_FRQ ≤ 0.7	3999366	4178261	4116117	3204722	3080262	2876657	**3575897**	**69.3**
2	Total DP ≥ 60	457791	602056	651579	471732	467012	428680	**513141**	**9.9**
3	Total DP ≤ 400	2809	1594	1278	943	765	1030	**1403**	**0.03**
4	per samp DP ≥ 30	201763	150667	161703	136596	129643	115373	**149290**	**2.9**
5	GQ ≥ 99	16540	17379	16993	18993	14176	18598	**17113**	**0.3**
6	DBB ≤ 100	1568	43	51	65	56	31	**302**	**0.01**
	Original SNPs	5645999	6051194	5974602	4708192	4467496	4132498	**5163330**	
	Filtered	4679837	4950000	4947721	3833051	3691914	3440369	**4257148**	**82.5**
	Remaining	966162	1101194	1026881	875141	775582	692129	**906181**	

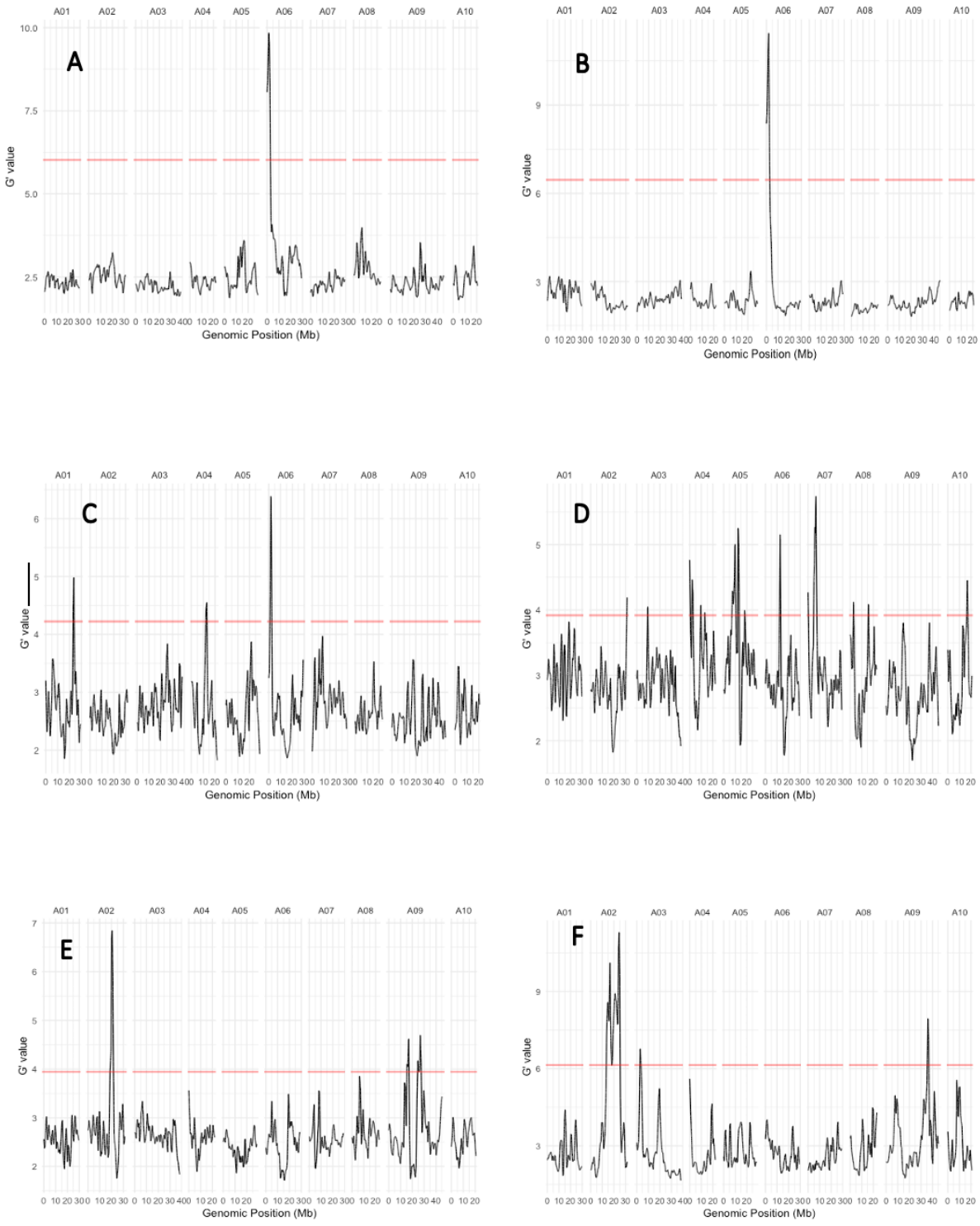


Figure 2.18. WGS-BSA mapping of resistance to two *Albugo candida* race 2 isolates (AcBj12 and AcBjDC) in five *Brassica rapa* genebank accessions. **A)** shows EH_25 exclusively mapped with Donskaja-virulent isolate AcBjDC. Whilst **B-F** show all five accessions (B = EH_25, C = EH_5, D = EH_47, E = EH_61, F = EH_5) with mapped resistance to Donskaja-avirulent isolate AcBj12. Resistant and susceptible phenotype bulks were produced at 14 dpi. QTL is displayed using the –log10 (p-value) which is derived from the G' statistic. The red line indicates the genome-wide false discovery rate (FDR) of 0.001.

Table 2.6. Summary of significant QTL regions for resistance to *Albugo candida* race 2 in five *Brassica rapa* genebank accessions (FDR of 0.001). Calculated using WGS-BSA and the QTLseqr pipline and the G' approach. Table includes chromosome start and end positions for each QTL, plus the number of SNPs each contains. A mean *p*-value score is given per QTL. The mean Q-value is the average adjusted *p*-value in the region. EH_25 is the only accession to be inoculated with both AcBj12 and AcBjDC.

Pop ID	QTL ID	Inoculum	Chr.	Start	End	Length	nSNPs	posPeak DeltaSNP	Max Gprime	posMax Gprime	meanPval	meanQval	NLR no. in QTL
5	EH_5.A01	AcBj12	1	23042028	23916720	874692	4738	23582433	5.0	23582433	1.69E-05	0.0015	2
61	EH_61.A02	AcBj12	2	18670297	21646283	2975986	9226	20643769	6.8	20625396	1.41E-05	0.0008	0
95	EH_95.A02.1	AcBj12	2	13506392	18172743	4666351	17239	18172743	10.1	16702949	7.37E-05	0.0018	20
95	EH_95.A02.2	AcBj12	2	18188233	25414269	7226036	22426	19645629	11.3	24564074	2.26E-05	0.0006	2
95	EH_95.A03	AcBj12	3	2803749	3747166	943417	2591	2803749	6.8	3037302	2.51E-04	0.0060	0
5	EH_5.A04	AcBj12	4	12167391	13148321	980930	5532	12334533	4.5	13019461	3.00E-05	0.0026	2
47	EH_47.A05.1	AcBj12	5	7122562	10560738	3438176	11325	10560738	5.0	9793968	1.98E-05	0.0009	0
47	EH_47.A05.2	AcBj12	5	12155492	13426805	1271313	2468	12897772	5.2	12485401	4.29E-06	0.0002	0
47	EH_47.A05.3	AcBj12	5	18194528	18445118	250590	263	18245030	4.0	18256222	2.45E-04	0.0081	8
5	EH_5.A06	AcBj12	6	948017	2715701	1767684	9841	948017	6.4	1834178	4.24E-06	0.0004	8
25	EH_25_12.A06	AcBj12	6	12787	2945454	2932667	9979	1378963	11.4	1834238	3.08E-06	0.0004	8
25	EH_25_DC.A06	AcBjDC	6	15873	2720230	2704357	7632	1383848	9.8	1383848	2.77E-06	0.0004	0
47	EH_47.A07	AcBj12	7	4774432	7584393	2809961	5193	6817217	5.7	6817217	9.45E-06	0.0004	0
61	EH_61.A09.1	AcBj12	9	15438675	17164935	1726260	6622	16931552	4.6	16931552	5.01E-05	0.0032	1
61	EH_61.A09.2	AcBj12	9	24490331	27607522	3117191	11925	24706707	4.7	26813210	1.13E-04	0.0056	0
95	EH_95.A09	AcBj12	9	35474968	36853630	1378662	4903	36689196	7.9	35982177	6.28E-05	0.0017	0

2.5.3 Positions of WGS-BSA QTLs relative to GBS raSNPs. For a comparison of results from the two resistance mapping approaches, raSNPs (conserved dominant) from GBS-mapping were overlayed with QTLs from WGS-BSA mapping (Figure 2.19). Generally, there is little correlation between the two datasets as only three raSNPs (EH_25 on A06) out of 22 (from the five accessions used for WGS-BSA mapping) are positioned inside the equivalent accessions QTLs. The remainder seem to have a random association and are frequently found on different chromosomes to their relevant accession QTLs. However, the location of all three EH_25 raSNPs (AcBj12) inside the EH_25_12.A06 QTL is a strong indication that the same WRR locus is being successfully mapped using both methods. Interestingly, EH_25 was represented for GBS-mapping by the most resistant individuals (49) relative to the other accessions. It was also the only accession to produce a single major-effect QTL from WGS-BSA mapping.

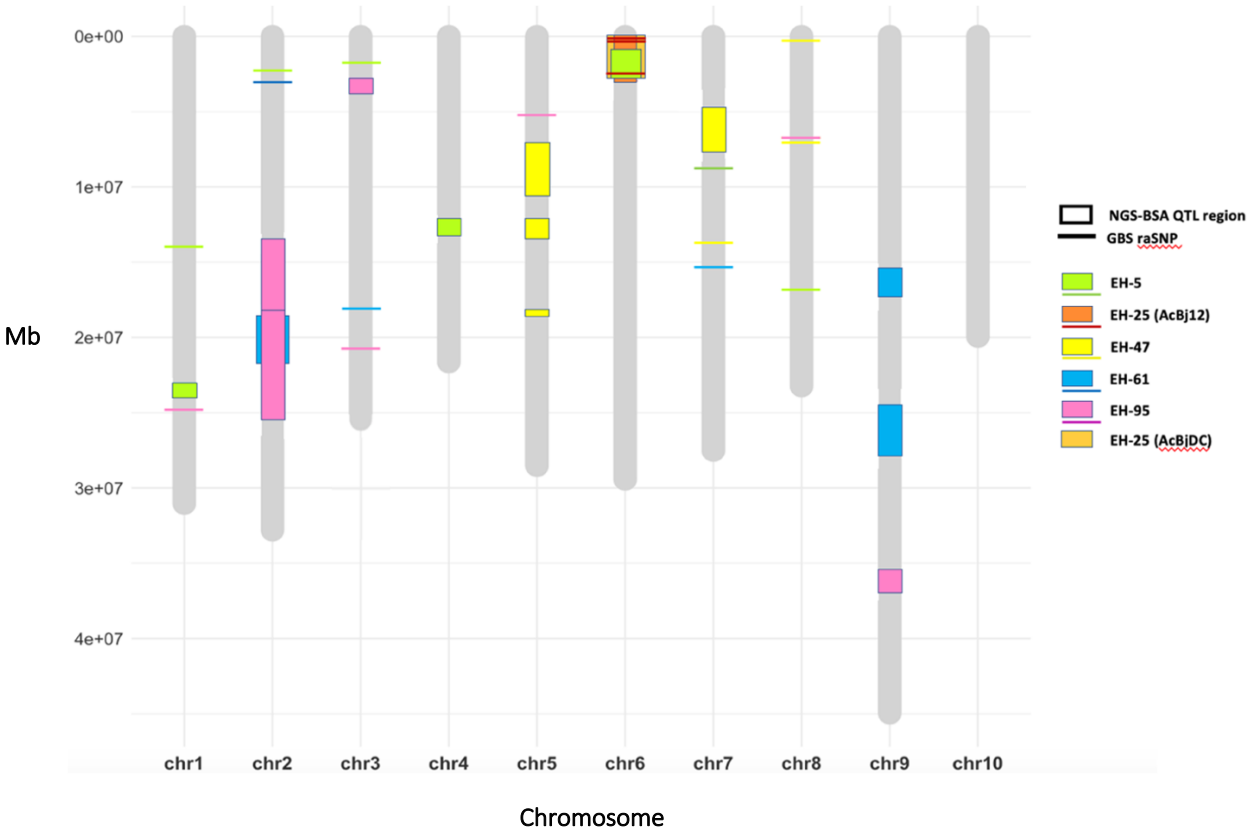


Figure 2.19. Physical map of the *Brassica rapa* reference assembly, showing the positions of GBS-mapped raSNPs (dominant conservative) (coloured lines) relative to QTLs (coloured blocks) mapped with WGS-BSA methods, for five genebank accessions. EH_25 has been represented twice in the WGS-BSA experiment, once tested with AcBj12 and once with AcBjDC. The only example where approximately the same WRR locus is identified using both methods is at the top of chr6, where three EH_25 (AcBj12) raSNPs are positioned inside the EH_25_12.A06 QTL.

2.5.4 Positions of WGS-BSA QTLs relative to *R*-genes. QTLs mapped using WGS-BSA were assessed relative to positions of annotated major *R*-genes classes in the *B. rapa* IVFCAASv3 reference genome (see Figure 2.20 and Table 2.7). Using the full suite of *R*-gene families outputted from DRAGO 2, all 16 QTLs from five accessions span regions that contain candidate resistance genes. Seven out of the 16 QTLs span regions containing genes from the major classes of cytoplasmic TNL and CNL receptors. The key regions of BraA02 and BraA06 both span multiple *R*-genes (eight and 13, respectively). BraA02 region contains a single TNL gene (BraA02g031160) and two RLKs (BraA02g025430 + BraA02g026280), whereas the BraA06 region contains a cluster of five CNL genes (BraA06g002390, BraA06g003120, BraA06g003420, BraA06g003430 + BraA06g003770) and four RLKs (BraA06g002310, BraA06g002330, BraA06g002340, BraA06g002360). Other QTLs containing TNL and CNL classes of *R*-gene include EH_95.A03 which covers the TNL BraA03g008470 on chromosome 3, and EH_47.A05.1 which covers the CNL BraA05g013230.

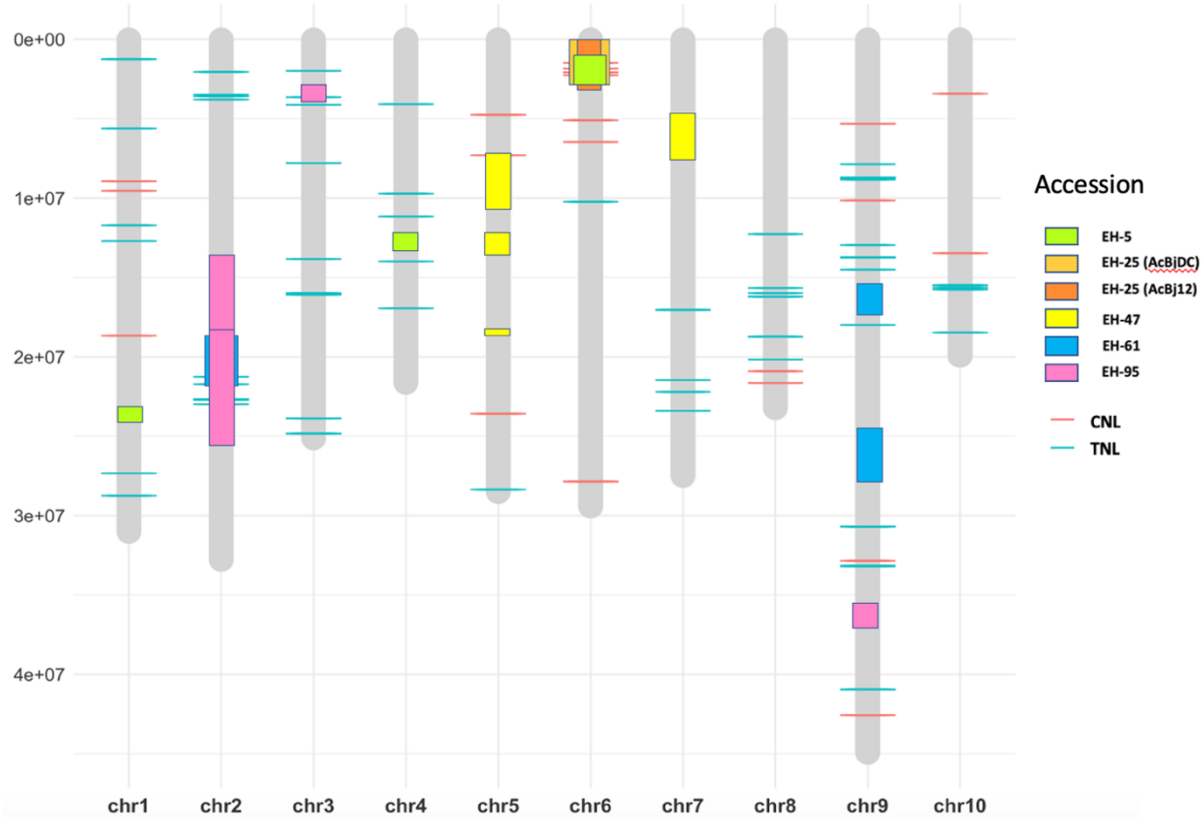


Figure 2.20. Physical map of the *Brassica rapa* reference assembly showing positions of WGS-BSA QTLs (coloured blocks) relative to major *R*-gene classes of CNL and TNL receptors, for five genebank accessions.

Table 2.7. Tabulated positions of all *Brassica rapa R*-genes (annotated using DRAGO 2 hidden Markov model software) physically located within QTL boundaries. This is a collection of WGS-BSA mapping results from five genebank accessions.

Chr	Accession QTL	QTL pos.	*R*-gene	Start pos.	Gene ID	Length (bp)
1	EH_5.A01	23042028				
			RLP	23250564	BraA01g033950	2522
			CK	23460000	BraA01g034200	2368
	EH_5.A01	23916720				
2	EH_95.A02.1	13506392				
			CK	14916326	BraA02g024950	4024
			RLK	15276025	BraA02g025430	6251
			CK	15307031	BraA02g025490	2027
			RLK	15736348	BraA02g026280	3599
			CK	16411969	BraA02g027010	2003
			CK	17945401	BraA02g028670	3580
	EH_95.A02.1	18172743				
	EH_95.A02.2	18188233				
	EH_61.A02	18670297				
			RLK	20707211	BraA02g030530	5547
			TNL	21266944	BraA02g031160	5931
			RLK	21489193	BraA02g031450	6182
			RLP	21553166	BraA02g031530	3170
			RLP	21560904	BraA02g031540	3014
			T	21639494	BraA02g031600	976
			NL	21640978	BraA02g031610	2170
			T	21645948	BraA02g031620	539
	EH_61.A02	21646283				
			T	21668513	BraA02g031630	608
			NL	21671827	BraA02g031640	5572
			T	21682868	BraA02g031650	1616
			CTNL	21690398	BraA02g031660	4277
			TNL	21721969	BraA02g031710	8231
			CN	22006775	BraA02g032040	2195
			RLK	22169791	BraA02g032180	3178
			RLK	22615923	BraA02g032770	3656
			TNL	22679826	BraA02g032840	4979
			CT	22689939	BraA02g032850	467
			N	22690837	BraA02g032860	1420
			NL	22697547	BraA02g032880	4650
			TNL	22708626	BraA02g032890	4225
			TNL	22727424	BraA02g032920	4276

Table 2.7. Continued

2			RLP	22864900	BraA02g033130	1654
			RLP	22983576	BraA02g033280	1792
			TNL	22988009	BraA02g033300	3369
			N	23873443	BraA02g034580	3204
			CK	23988923	BraA02g034690	2518
			RLP	24298773	BraA02g035120	2979
			RLK	24604525	BraA02g035420	2831
			CT	25028210	BraA02g036020	560
			T	25058968	BraA02g036040	625
			T	25072116	BraA02g036050	580
			RLP	25388090	BraA02g036490	8499
	EH_95.A02.2	25414269				
3	EH_95.A03	2803749				
			RLK	3108539	BraA03g007260	3479
			RLP	3248782	BraA03g007610	746
			RLP	3250839	BraA03g007620	728
			TNL	3638481	BraA03g008470	3484
			T	3665776	BraA03g008540	1642
	EH_95.A03	3747166				
4	EH_5.A04	12167391				
4			RLP	12781016	BraA04g016800	587
4			RLK	13063146	BraA04g017130	2533
4	EH_5.A04	13148321				
5	EH_47.A05.1	7122562				
			CNL	7301160	BraA05g013230	2582
			CK	9952022	BraA05g016740	1478
			RLK	10150191	BraA05g017000	4867
			CK	10369339	BraA05g017290	4474
	EH_47.A05.1	10560738				
	EH_47.A05.2	12155492				
			RLP	12200454	BraA05g019240	2585
			RLP	12228174	BraA05g019250	3047
			RLP	12583786	BraA05g019590	1811
	EH_47.A05.2	13426805				
	EH_47.A05.3	18194528				
			RLP	18277084	BraA05g024400	1442
	EH_47.A05.3	18445118				

Table 2.7. Continued

6	EH_25_12.A06	12787				
	EH_25_DC.A06	15873				
			RLK	729161	BraA06g001180	2827
	EH_5.A06	948017				
			T	1053954	BraA06g001690	693
			N	1143970	BraA06g001890	1308
			RLK	1420096	BraA06g002310	5439
			RLK	1431545	BraA06g002330	3817
			RLK	1437114	BraA06g002340	4678
			RLK	1453223	BraA06g002360	4073
			CNL	1477751	BraA06g002390	2540
			N	1688841	BraA06g002870	4841
			CNL	1831588	BraA06g003120	2558
			CNL	2067355	BraA06g003420	2531
			CNL	2072095	BraA06g003430	2555
			N	2214326	BraA06g003690	650
			CNL	2246940	BraA06g003770	2564
	EH_5.A06	2715701				
	EH_25_DC.A06	2720230				
	EH_25_12.A06	2945454				
7	EH_47.A07	4774432				
			RLK	5131735	BraA07g006070	3325
			RLK	6949607	BraA07g006770	2100
	EH_47.A07	7584393				
9	EH_61.A09.1	15438675				
			T	16021466	BraA09g023950	1034
	EH_61.A09.1	17164935				
	EH_61.A09.2	24490331				
			CK	27053695	BraA09g034380	3416
	EH_61.A09.2	27607522				
	EH_95.A09	35474968				
			RLK	35594894	BraA09g046870	2819
			RLK	35714543	BraA09g047080	3292
			RLK	35785877	BraA09g047260	3334
			CK	36397077	BraA09g048400	1055
			RLK	36422709	BraA09g048440	2523
	EH_95.A09	36853630				

2.5.5 An automated assessment of candidate genes within mapped QTL intervals (autoSNP).

A major benefit of using WGS data compared to other marker techniques is that the dataset used to identify QTLs can then be screened at the sequence level to look for candidate mutations. This is achieved by checking genome assemblies within QTLs and screening gene coding sequences (CDS) for mutations with allele frequencies that distinguish phenotype bulks. Here for example, where major effect dominant *R*-genes are thought to be the main precursors to WRR, SNPs that are homozygous in the susceptible bulk and heterozygous in the resistant should be considered. This could also be applied to look for recessive resistance by reversing this pattern (homozygous resistant vs heterozygous susceptible).

An automated method (termed autoSNP) was used in this study for filtering available WGS SNP datasets to identify genes within mapped QTLs that contain dominant associated mutations. Firstly, SNPs within the relevant accession QTLs were selected for. This list was then filtered to identify those located within CDS regions of genes. SNPs with coverage < 30x were removed from the dataset. Remaining SNPs were compared between phenotype bulks using vcftools (v0.1.16, (Danecek et al., 2011)) to identify shared positions where allele frequencies (AF) were suitably distinct. Parameters determining selection were: homozygous SNPs in the susceptible bulk with AF < 0.1 or AF > 0.9 versus heterozygous SNPs in the resistant bulk with AF > 0.4 & AF < 0.6. A subset of the final markers was checked manually in Integrated Genome Viewer (IGV) to confirm they met these criteria. Genes containing markers identified by the autoSNP process were blasted in Phytosome (Goodstein *et al.*, 2012) against the *B. rapa* reference genome FPsc v1.3 to obtain gene descriptions (Table 2.8). Amongst all accessions, 14 out of 16 QTLs contain genes identified using the autoSNP method. The total number of genes in the *B. rapa* IVFCAASv3 reference annotation file (.gff) contained within all QTLs is 4884. The total number of genes identified using autoSNP within all QTLs was 272, which accounts for 5.75% of the total gene count. Of these genes, those described (using a combination of Phytosome and DRAGO 2 *R*-gene annotations) as having a potential role in disease resistance have been presented in Table 2.8. These are each presented along with an available marker selected by autoSNP.

Table 2.8. Candidate genes identified within QTL loci of *Brassica rapa* using the autoSNP method. Genes were selected on the basis of containing a dominantly associated SNP (homozygous susceptible AF < 0.1 or AF > 0.9 vs heterozygous resistant AF > 0.4 & AF < 0.6) within CDS gene regions. Those with descriptions related to *R*-genes (using a combination of Phytosome and DRAGO 2 annotation) have been listed. This method provides an assessment of all genes contained within mapped QTLs.

Accession	Chr	QTL	No. genes in QTL	No. *R*-genes	*R*-gene	Marker pos.	Description
EH_5	1	EH_5.A01	122	10	BraA01g033920	23226809	Transcription initiation factor
EH_61	2	EH_61.A02	240	14	BraA02g030540	20722322	Cyclin dependent kinase
EH_95	2	EH_95.A02.1	582	21	BraA02g025430	15277340	RLK
		EH_95.A02.2	759	35	BraA02g031160	21267472	TNL
					BraA02g031540	21562914	RLP
					BraA02g032880	22700315	NL
					BraA02g032890	22710556	TNL
	3	EH_95.A03	228	5	None detected		
EH_5	4	EH_5.A04	120	11	BraA04g016530	12562680	Kinase interacting protein
					BraA04g016540	12567323	Containing Protein
EH_47	5	EH_47.A05.1	446	9	BraA05g013580	7572697	Transcription factor
		EH_47.A05.2	132	0	None detected		
		EH_47.A05.3	25	0	None detected		
EH_25	6	EH_25_12.A06	489	37	BraA06g001440	880367	RNA-Binding Protein
					BraA06g001980	1186971	D-mannose kinase domain
					BraA06g002480	1508992	Lectin binding protein
					BraA06g003420	2069204	CNL
EH_25	6	EH_25_DC.A06	461	47	BraA06g002300	1416918	RLK
					BraA06g003420	2069204	CNL
EH_5	6	EH_5.A06	309	50	BraA06g002300	1417199	RLK
					BraA06g002360	1454484	RLK
					BraA06g003420	2069204	CNL
EH_47	7	EH_47.A07	153	6	BraA07g006070	5133150	RLK
					BraA07g006770	6951408	RLK
EH_61	9	EH_61.A09.1	335	9	BraA09g024950	16752708	Salt stress response kinase
					BraA09g025250	16958583	Protein tyrosine kinase
		EH_61.A09.2	219	3	None detected		
EH_95	9	EH_95.A09	264	24	BraA09g048400	36397457	CK
					BraA09g048440	36422792	RLK

2.5.6 Secondary screening of the BraA02 and BraA06 region. Repeated mapping of the BraA02 region on A02 using EH_61 and EH_95, as well as the BraA06 region with EH_25 (AcBj12 and AcBjDC) and EH_5 provides an opportunity for screening *R*-genes in these regions for conserved functional resistance alleles. The eight *R*-genes in BraA02 and 13 in BraA06 were checked for those showing a conserved resistance allele that matches a pattern of homozygosity in all susceptible bulks vs heterozygosity in all resistant bulks. Due to subtle variations in allele frequencies apparent in WGS bulked samples, only distinct SNPs matching these criteria were considered. Inspection of these regions was performed both visually, using the superimposed resistant and susceptible .bam file sequences of the relevant accessions in IGV, and using outputs from the autoSNP analysis listed in Table 2.8.

The only *R*-gene example found to fit the expected pattern was the CNL gene BraA06g003420 located in the BraA06 region. Other examples in this cluster either contained visually ambiguous allele ratios, were not conserved between accessions, or lacked partial sequence coverage and so could not be confirmed. BraA06g003420 contains three distinct heterozygous SNPs at positions 2069204, 2069428 and 2069463 in the resistant bulk and homozygous in the susceptible. These SNPs are fully conserved in the same expected ratios between the three experiments of EH_25 (AcBj12 and AcBjDC) and EH_5 (Figure 2.21). To determine the effect these mutations may have on protein translation and structure, the mutations were checked for effect on amino acid sequence. It was found that two of the three SNPs (positions 2069204 and 2069428) result in non-synonymous amino acid substitutions. The A > T at position 2069204 results in a lysine to threonine substitution, whilst the G > A at 2069428 results in an alanine to threonine substitution. The remaining SNP (2069463) produces a synonymous amino acid substitution and is unlikely to affect protein structure or function.

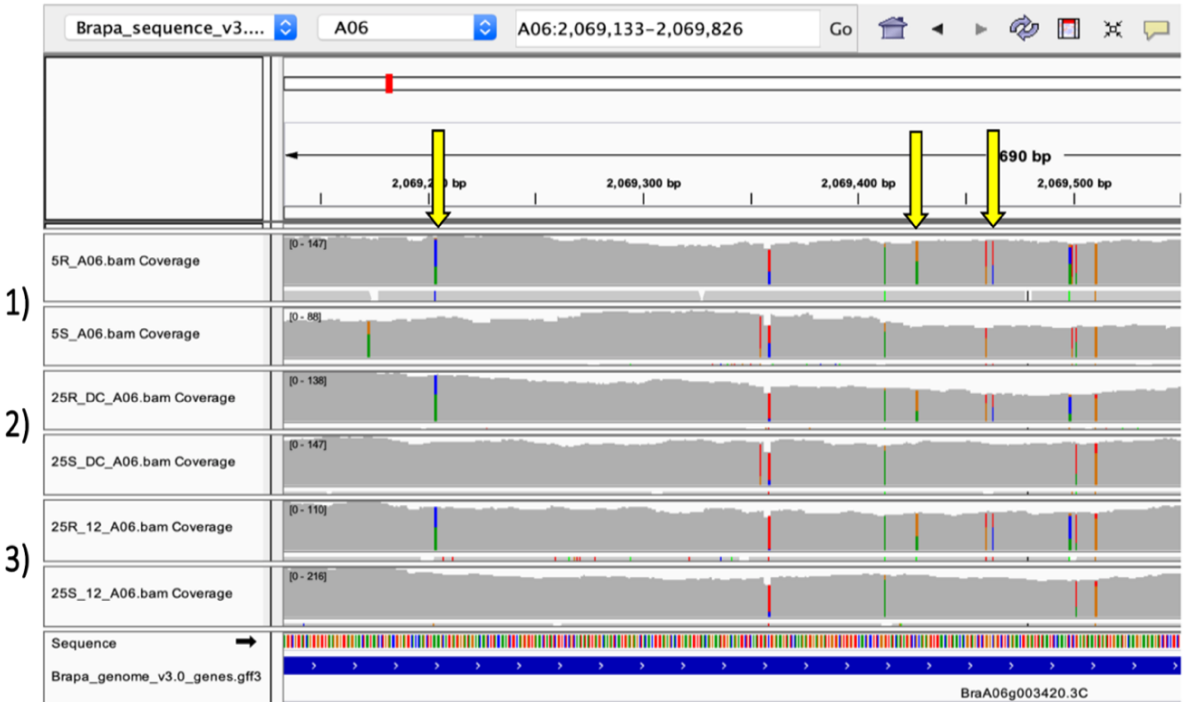


Figure 2.21. Trace files of resistant and susceptible WGS bulks (top and bottom respectively for each of 1, 2 and 3) visualised in IGV, showing the C terminus of the coiled-coil CNL candidate BraA06g003420 for **1)** EH-5; **2)** EH-25, tested with AcBjDC; **3)** EH-25, tested with AcBj12. The bottom bar denotes an annotated gene track from the *B. rapa* reference genome (v3.0). SNPs that distinguish phenotypes and are conserved between accessions are denoted by the yellow arrows at positions 2069204, 2069428 and 2069463. This would be the expected pattern for any causal dominant candidate gene.

2.6 DISCUSSION

2.6.1 Discussion of GBS-mapping method

The rapid evolution of plant pathogens relative to their hosts allows them to frequently overcome immune defences under the control of a single pathogen resistance gene (*R*-gene). Subsequently, there is a need to continuously identify new *R*-genes, that when combined together in a cultivar will provide durable protection of the crop. A major aim of this thesis was to trial new methods for rapidly mapping WRR genes directly from a range of genetically diverse genebank material. Two distinct approaches were deployed to achieve this: GBS-mapping and WGS-BSA mapping. Here, each method is considered accordingly, with results from the initial GBS-mapping experiment informing the decision to then proceed with WGS-BSA methods. In turn, WGS-BSA results were then used to validate the accuracy of the GBS-mapping data.

Identification of 36 resistance-associated SNPs (raSNPs) from GBS-mapping of eight *B. rapa* accessions was an insightful first attempt at mapping resistance loci directly from segregating genebank material. This was achieved using GBS of a susceptible bulk and multiple resistant individuals per accession, and extensive filtering of genotype data to select SNPs with strong association to phenotype (raSNPs). Considering the novelty of the experimental method, as well as the potentially heterozygous nature of the germplasm, there were no prior expectations regarding the generation of raSNPs. The output of 36 candidate raSNPs that passed the highly stringent filtering parameters was therefore initially deemed successful. Furthermore, that such a large proportion (38%) of these markers are found within 200 kb of known *R*-genes was a promising indication that this marker-trait association method was effective in identifying resistance loci. When assessing the results however, it became apparent there were potential issues with the data quality. For example, prior expectations were that if raSNPs were generated at all, that multiple associated markers would coalesce around a given resistance locus. This was not the case and raSNPs were almost exclusively found in isolation, with few close proximity markers (< 1 kb) meeting the filtering criteria for association to phenotype. Markers proximate to raSNPs were frequently not 'borderline' candidates either but actually showed very little or highly variable associations. When considering a single marker in this manner with no proximate support, there is no local 'replication' (and

therefore confidence) of association to suggest a candidate raSNP is not a false positive. Arguably the lack of local association seen in this dataset could be due to very high rate of recombination in these lines. However, assessment of LD decay (Figure 2.11) using the GBS-subset suggests that for so many raSNPs, this would be improbable across such small distances.

The main factor determining the number of outputted raSNPs was the number of resistant individuals included per accession for sequencing (Figure 2.4). Lines represented by few resistant individuals (< 25) typically outputted more raSNPs, whilst those represented by lots of individuals (> 40) typically produced the fewest. Essentially, the greater the number of resistant individuals included, the greater the opportunity for disagreeing genotype calls between phenotypes. Accessions represented by higher numbers therefore have a higher stringency than accessions represented by fewer individuals and when screening many millions of markers it appears that those represented by fewer individuals will generate raSNPs just by chance.

The unexpectedly high genotype error rate of 13.6% between the two biological control susceptible bulks was a further issue impacting the accuracy of predicted raSNPs. This was not foreseen in the design of the experiment and is likely a result of 'grey area' allele frequencies (between a homozygous (0 or 1) and heterozygous state (0.5)), that deviate sufficiently between susceptible bulks to produce different genotype calls. This will have resulted in a proportion of raSNPs being unnecessarily excluded (where the susceptible bulk genotypes do not match). True homozygosity in the susceptible bulks (AF = 0) was also rare and will have effected filtering when using conservative parameters.

With these factors impacting the potential validity of the GBS-mapping results, it was decided to pursue a more conventional mapping approach using WGS-BSA and the QTLseqr pipeline. With hindsight this method proved to be much more creditable for reasons that will later be discussed. The more statistically rigorous WGS-BSA approach highlighted the extent of the GBS-mapping experimental design flaws by showing the wide discord between positions of GBS-mapped raSNPs relative to WGS-BSA mapped QTLs (Figure 2.19).

From the five *B. rapa* accessions used for WGS-BSA, only the three GBS-mapping raSNPs from EH_25 were successfully positioned within the relative accessions QTL. This wild species was represented by the most resistant individuals (49) out of any accession

and tentatively indicates the numbers required to make accurate associations. As the only accession to map a single major effect resistance locus (as demonstrated by QTLs EH_25_12.A06 and EH_25_DC.A06), EH_25 demonstrates that the GBS-mapping method seemingly breaks down where multiple loci are involved. The identified EH_25 raSNPs were a satisfying yet solitary example that it is possible using the GBS-mapping method to produce true raSNPs. Considering however that GBS-mapped raSNPs for other accessions were mostly not even on the same chromosome as their constituently mapped QTLs, it suggests that the approach is far from efficient, with only a very limited output relative to the time and expenses invested.

2.6.2 Discussion of WGS-BSA method

Of the two methods, WGS-BSA mapping proved far superior, identifying sixteen distinct QTLs across five *B. rapa* accessions that likely contain major effect WRR genes. These were defined by high-resolution QTLs (averaging 2.44 Mb), with all sixteen spanning some form of potential *R*-gene. Seven of these QTLs contained the major classes of TNL and CNL resistance genes. WGS-BSA mapping was achieved rapidly, in a span of approximately five months from screening germplasm to mapped outputs, and at minimal cost requiring WGS of only two bulks per accession. The relative simplicity of processing two tissue bulks per accession in turn enabled screening and mapping of multiple lines in parallel to identify a range of WRR loci within a diverse range of *B. rapa* germplasm.

With costs of next generation sequencing (NGS) continuing to fall, WGS became a viable option for this project, allowing complete resequencing of *B. rapa* tissue bulks. Low prices meant that enhanced 90x sequencing coverage could be applied, accounting for heterozygosity in the genebank accessions and maximising capture of tens of thousands of high-quality SNPs. Mapping of WRR was achieved using the recently developed WGS-BSA (QTL-seq) R package QTLseqr (Mansfeld & Grumet, 2018), which made it possible to progress rapidly in identifying QTLs. Use of WGS datasets provided several advantages over other marker technologies. For example, unlike those that reduce the genome complexity prior to sequencing such as GBS or RAD (Andrews *et al*., 2016), WGS datasets provide a near-complete view of sequence variation and highlights a set of potential markers across defined QTL regions. Also, relative to bait design technologies such as

RenSeq (Jupe *et al*., 2013), which are limited to the selection of genes chosen for the bait, WGS requires no *a priori* knowledge about the types of genes responsible for phenotype.

QTL-seq has been used to successfully map traits in other crop species including; seed weight in chickpeas (Das *et al*., 2014; Singh *et al*., 2016), early flowering in cucumber (Lu *et al.*, 2014), rust and late leaf spot resistance in groundnut (Pandey *et al*., 2017), a cytosolic component of photosynthesis efficiency in potato (Kaminski *et al*., 2015), tomato fruit weight (Illa-Berenguer *et al*., 2015), cold tolerance in rice (Luo *et al*., 2018), resistance to crown rot in squash (Ramos *et al*., 2020), plant height in soybean (X. Zhang *et al*., 2018), flowering time in chickpea (Srivastava *et al*., 2017), fusarium wilt resistance in watermelon (Branham *et al*., 2018) and heat tolerance loci in broccoli (Branham & Farnham, 2019). Of note, these studies all utilised RILs and F_2 recombinants from biparental crosses (where the parents have previously been fixed in a homozygous state) to produce the phenotypic extremes required for tissue bulks. The approach undertaken in this study therefore represents the first known example of QTL-seq methods being applied directly to undeveloped genebank material, utilising the phenotypic variation found with a single packet of seed to generate sufficient mapping bulks. Importantly, by using this approach a greater range of genebank material can be accessed that includes genetically diverse landraces and wild species. The outcrossing nature of *B. rapa* is particularly well suited to this, where the maintenance of genebank accessions through controlled intercrossing within small accessions preserves the genetic variation required for mapping. This suggests that other outcrossing species may be equally receptive to this approach.

The majority of marker-trait analysis performed in plants, including previous WRR mapping efforts, have been undertaken using linkage mapping (LM). These experiments require construction of a mapping population from crossing two genetically fixed parents. This typically requires multiple generations of inbreeding, that depending on the species can take years to produce. Maintaining the large number of individuals required for an effective analysis is also expensive, time-consuming and labour intensive. Furthermore, mapping populations are derived from the genetic hybridization of two parental genotypes with an alternative trait of interest, which limits variation to that contained within the two parents. This represents a tiny amount of the variation contained within *ex situ* genebanks and typical germplasm collections. When considering the significant

expenditures in time, cost and resources required by using LM methods, strategies that utilize genebank resources directly are comparably highly effective.

A key WGS-BSA result here was the mapping of the chromosome 6 region BraA06, using the wild species accession EH_25 (tested independently with both AcBj12 and AcBjDC) and the landrace EH_5. These highly significant QTLs, particularly in EH_25 indicate a single major effect resistance locus, confined to a 2.93 Mb window. This result was particularly distinguished by repeatedly mapping the BraA06 region with a total of four independent sources, including: EH_25 GBS-mapping raSNPs, the EH_25 AcBj12 QTL (EH_25_12.A06), the EH_25 AcBjDC QTL (EH_25_DC.A06) and the EH_5 AcBj12 QTL (EH_5.A06). Firstly, this provided validation of the EH_25 raSNPs and to a limited extent the GBS-mapping method. Secondly, mapping the same locus using both race 2 isolates strongly implies that a single major resistance gene controls resistance to both isolates. Finally, mapping the same locus using the advanced cultivar EH_5 is suggestive that this resistance allele is conserved between both accessions. This is supported by their shared ancestry as demonstrated from the neighbor joining tree in Figure 2.12, as well as their relatively close geographic origins (EH_5 from Italy and EH_25 from Algeria).

Annotated *R*-genes within the cumulative BraA06 window highlight a number of candidate WRR genes, including the largest cluster of coiled-coil NLRs (CNL) in the *B. rapa* genome. These five CNLs, five RLKs and three other *R*-genes in this region are all strong resistance candidates, with bulked WGS data only able to rule out BraA06g001690 as monomorphic. By using a combination of automatedly searching for highly associated SNPs (autoSNP), as well as visual assessment of sequence files in IGV, the CNL BraA06g003420 emerged as a strong candidate. The three non-synonymous SNPs displayed in Figure 2.21 represent the most distinct dominant allele frequencies (homozygous susceptible vs heterozygous resistant) of any SNPs in BraA06 *R*-genes. Conservation of these markers in the same distinct ratios in both accessions lends support to BraA06g003420 being a functional WRR gene.

The other major QTLs on chromosome 2 (EH_61.A02 in EH_61 and EH_95.A02.2 in EH_95) whose overlapping locus is termed here as BraA02, enabled a similar assessment for conserved *R*-gene SNPs in the region. However, alleles for the eight BraA02 *R*-genes do not appear to be conserved between the two accessions. The strength of SNP association in EH_61 is also not sufficient to distinguish any candidate genes for

this accession in the EH_61.A02 QTL. For EH_95, only the EH_95 QTL EH_95.A02.2 showed allelic associations strong enough to be detected with either the automated method or visually in IGV. The EH_95.A02.2 *R*-genes displaying particularly strong associations include: BraA02g032840 - TNL, BraA02g032850 - CT, BraA02g032860 - N, BraA02g032880 - NL and BraA02g032890 - TNL. This 28.8 kb region, which lies approximately between positions 22679826 and 22708626 has distinct SNP associations to phenotype relative to anything further out, making these *R*-genes of particular interest. Furthermore, the orthologue of the TNL gene BraA02g032840 is At5g46270, which has been identified as the WRR gene *WRR8* in the *Arabidopsis* background Sf-2 and confers resistance to the race 2 isolate Ac2v. Subsequent transformation of the *WRR8* allele into a susceptible *B. juncea* background however did not produce Ac2v resistance and it is speculated that *WRR8* resistance in *Arabidopsis* involves a guardee that is present in *Arabidopsis* but not in Brassica spp. However, *WRR8* remains untested in a *B. rapa* background.

Complimenting WGS-BSA mapping with *R*-gene annotation (DRAGO 2 software) and raSNP analysis proved to be an effective way to scrutinise mapped QTL regions for candidate genes. The autoSNP method for example was able to identify some of the most highly associated *R*-genes in these datasets, such as BraA02g032890 in EH_95 and BraA06g003420 in EH_5 and EH_25. However, scrutiny of QTLs regions using bulked WGS datasets is limited in terms of what can be inferred about candidate genes. This is because bulk samples represent a combined or 'averaged' recombination window where any polymorphism linked to the functional gene shows a degree of allelic bias towards phenotype. The extent of any marker association is directly proportional to distance from the functional gene, with those in close proximity showing higher associations. Whilst this regional association provides the necessary signal required to generate QTL intervals it hinders identification of any single candidate gene. The tendency of major *R*-gene families such as TNL and CNL receptor genes to cluster in the genome further hampers efforts to distinguish individual candidates.

Whilst the average resolution of QTLs mapped here is relatively high (average physical interval size of 2.44 Mb), it is not possible to refine them further without design of new experiments. Primarily, this would require screening individual recombinants using markers obtained from the *R*-genes located within mapped QTLs in a classic LM

experiment. Candidate genes would then be retained or excluded based on their segregation with phenotype. This highlights a major benefit of using WGS datasets as opposed to standard marker technologies, where the abundance of markers obtained (particularly those within *R*-genes) can then be utilised for fine-mapping experiments. Effective LM depends on not only the abundance of available markers however but also on the availability of recombinant individuals (Cantor, 2018). This is potentially a limiting factor when using packets of genebank seed, which here weigh approximately 0.5 grams and supplying approximately 200 – 400 *Brassica* seeds. Where stocks are limited crosses would need to be made to generate necessary F_2 segregants. In this instance, the broadly susceptible and rapid cycling *B. rapa* line R-O-18 would provide a suitable parent for crossing resistant material to. An alternative approach to fine mapping would be to clone candidate WRR alleles into a susceptible background for functional testing. Availability of the broadly susceptible *Arabidopsis* background MAGIC-MAGIC, as well as relatively simplistic modern cloning systems such as Golden Gate (Engler *et al*., 2008) and USER protocols (Geu-Flores *et al*., 2007) makes direct testing of alleles a tempting option here over further fine mapping.

Obtaining the F_1 resistance phenotype for screened genebank accessions would be recommended for future work, with the suggestion being to generate them alongside the WGS-BSA experiment. The QTLseqr method (Mansfeld & Grumet, 2018) does not differentiate between dominant and recessive resistance and knowledge of this would determine the relevant raSNP ratios to screen for, as well as the classes of genes assessed within QTLs. As the first test of this method directly applied to genebank material, analysis of QTL regions here focused on major effect dominant resistance genes, though it is possible that some of the mapped WRR loci here could be recessive. As recessive resistances are a vital source of both broad spectrum and durable resistance it is important that future work using this method accounts for both types. If prioritising recessive resistance, one would need to perform initial crosses (using a broadly susceptible parent) with a selection of genebank accessions to identify recessive resistance from F_1 progeny, selecting only those with a susceptible phenotype for bulk production and mapping.

The use of the *BjuWRR1* resistance breaking isolate AcBjDC here for mapping WRR in wild species accession EH_25 provides an important demonstration of employing a

virulent isolate to identify the next 'best' resistance genes. As recently demonstrated by Dev *et al*. (2020), isolates have already been identified in India that can overcome *BjuWRR1* resistance. If *BjuWRR1* is to contribute to durable white rust resistance in Indian oilseed crops, then it is imperative that new resistance genes controlling Donskaja-virulent pathotypes are identified and combined alongside *BjuWRR1* for targeted deployment within commercial *B. juncea* cultivars. This can only be achieved using resistance breaking isolates to screen material with. Consequently, the isolation of the EH_25_DC.A06 *R*-allele for deployment alongside *BjuWRR1* in *B. juncea* crop varieties would pre-emptively provide targeted resistance against emerging virulent pathotypes. Alternatively, introgression of *BjuWRR1* by itself would likely squander the potential benefits of the *R*-gene by increasing the prevalence of adapted pathotypes.

To conclude, the application of WGS-BSA QTL-seq methods using undeveloped genebank germplasm was successful in revealing multiple WRR loci in a broad range of *B. rapa* accessions including advanced cultivars, a landrace and a wild species. The simplicity of the technique, as well as savings in time and expenses gained when compared to traditional LM methods enabled multiple accessions to be processed in parallel. This was achieved in a time span of approximately five months, a significant improvement on the five to six years required to develop a standard mapping population from a single bi-parental cross. Mapping in the diploid background of *B. rapa*, as opposed to the potentially recalcitrant tetraploid genome of *B. juncea*, facilitated identification of distinct WRR loci. This extends the potential benefits of these identified loci to a total of three globally important *Brassica* crop species, including *B. rapa* (AA), *B. juncea* (AABB) and *B. napus* (AACC). Importantly, these results demonstrate a more effective use of genebank resources for rapid mapping of WRR loci that may directly benefit sustainable oilseed mustard production in India.

Chapter 3

Investigating broad-spectrum white rust resistance in *Brassica oleracea* accession EBH527

3.1 INTRODUCTION

Recessive disease resistance plays an important role in crop protection as it can provide both broad-spectrum and durable disease control. The *Mlo* gene in barley is an exemplary demonstration of this, where a recessive mutant found seven decades ago has since maintained field resistance against all races of powdery mildew (Jørgensen, 1992). This first example of recessive resistance was described by Roland Biffen (1907) as an early application of genetics following rediscovery of Gregor Mendel's seminal work in the 1860s. Biffen investigated inheritance of a recessive allele for yellow rust resistance in wheat, which he then deployed in varieties that provided durable disease control. Similarly, Norman Borlaug deployed recessive stem rust resistance (*Sr2* locus) to breed durable wheat varieties as part of the Green Revolution in Mexico (Spielmeyer *et al*., 2003; Zeyen *et al*., 2002). The molecular mechanisms of susceptibility genes (*S*-genes) has emerged as a hot topic since cloning of the barley *mlo* gene (Büschges *et al*., 1997), and seminal research in *Arabidopsis thaliana* which led to *pmr6* encoding a pectate lysase-like gene for susceptibility to powdery mildew (Vogel *et al*., 2002), and *dmr6* encoding a novel oxygenase for susceptibility to downy mildew (Van Damme *et al*., 2008).

S-genes are now established as any host gene that is exploited by pathogens to facilitate the infection process. Since the earlier discoveries, numerous *S*-genes have been identified from a range of crop species, which are well documented in the review by van Schie & Takken (2014). Molecular mechanisms by which *S*-genes facilitate pathogenesis are diverse but share general features: 1) basic compatibility which the pathogen requires to initiate penetration; and 2) negative regulation of host immune signals to sustain growth and proliferation of the pathogen without detection by the host. Mutational disruption of *S*-genes in commercial crops is increasingly a key focus for resistance breeding and are ideal targets for loss-of-function mutation using gene-editing technology. However, as *S*-genes typically have an established evolutionary role in maintaining plant functions, loss-of-function mutants require careful monitoring for detrimental secondary effects on host fitness (*e.g.,* cell death or lesion mimic phenotypes).

Advances in genome editing technologies such as CRISPR-Cas are making it increasingly feasible to make precise alterations to DNA sequence, and as such are

heralded as a vital new tool within crop sciences. Application so far has enhanced both understanding and application of major agricultural traits such as yield, biotic and abiotic stress management (Yi Zhang *et al*., 2019). Most of the applications to date have been considered as 'proof-of-concept' within a particular plant species and have focused on knocking out one or several genes with known functions. Disease resistance has become an obvious targetable trait for the use of CRISPR loss-of-function experiments (Table 3.1). Firstly, because scientific understanding of specific pathosystems and their molecular mechanisms is sufficiently advanced to allow selection of the most likely gene candidates for assessment. Secondly, disease resistance is often controlled by the action of single, major-effect genes whose modification results in a distinct 'all-or-nothing' alteration to phenotype. This is generally easier to process than other traits such as abiotic stress tolerance, which are frequently under control of numerous genes, each providing incremental effects to phenotype. And thirdly, commercial advances in disease resistant crops are possible through targeted mutagenesis of *S*-genes, which is the most widely established function of CRISPR/Cas technology to date.

Previous PhD research (Fairhead, 2016) undertook mapping of a major effect recessive white rust resistance (WRR) locus in *Brassica oleracea*, screened using an isolate of *A. candida* race 9 (AcBoWells). The resistant parent EBH527 is a doubled haploid developed in an earlier DEFRA funded project (Holub, 2002) and was characterised for complete resistance to 18 isolates of *A. candida* race 9 collected from across the *B. oleracea* growing regions of the UK. This is the first known example of recessive WRR to be mapped in *B. oleracea* and is a potentially valuable source of broad-spectrum resistance for use as a future control of white rust in vegetable and oilseed *Brassica* crops.

A recombinant inbred mapping (RIL) population produced from the biparental cross of susceptible parent A12DH and EBH527 was used to map the recessive WRR locus. Genotyping-by-sequencing (GBS) marker technology (Elshire *et al*., 2011) enabled linkage mapping in the F_5 generation of a 1.2 Mb map interval on chromosome 2 designated Bo-ACA2 (Fairhead, 2016). An 18 kb interval (termed ACA2) was defined using a combination of markers generated from *R*-gene enrichment sequencing (RenSeq, (Jupe *et al*., 2013)) and reference-based markers developed from other genes within the Bo-ACA2 locus. The interval spans four candidate genes, including three that were previously shown via Sanger sequence to be monomorphic (Bo2g016510, Bo2g016500 and Bo2g016490) and

one (Bo2g016480) that contains a single SNP in the first exon that co-segregates with phenotype in all recombinant individuals tested.

Table 3.1. Susceptibility genes targeted using CRISPR/Cas gene-editing for engineered disease resistance. References were predominantly sourced from reviews by Zaidi *et al*. (2018) and Zhang *et al*. (2019). Abbreviations include; TG = contains transgene, TGF = transgene free, PM = powdery mildew.

Plant	*S* gene target	System	Targeted pathogen/disease	Result	Ref
Rice	*OsERF922*	TGF	Rice blast	Enhanced rice blast resistance	(F. Wang et al., 2016)
Tomato	*SlMlo1*	TGF	PM	Resistance to PM	(Nekrasov *et al*., 2017)
Arabidopsis	*eIF4E*	TGF	TuMV	Resistance to TuMV	(Pyott *et al*., 2016)
Cucumber	*eIF4E*	TGF	CVYV, ZYMV and PRSMV (virus)	Broad resistance to specified viruses	(Chandrasekaran *et al*., 2016)
Grape	*MLO-7*	TGF	PM	Enhanced PM resistance	(Malnoy *et al*., 2016)
Apple	*DIPM-1, DIPM-2,* and *DIPM-4*	TGF	Fire blight (*Erwinia amylovora*)	Enhanced Fire blight resistance	(Malnoy *et al*., 2016)
Tomato	*SlJAZ2*	TG	Bacterial spec	Bacterial spec resistance	(Ortigosa *et al*., 2019)
Grapefruit	*CsLOB1* promoter	TG	Citrus canker	Alleviated citrus canker	(Jia *et al*., 2016)
Grapefruit	*CsLOB1*	TG	Citrus canker	Citrus canker resistance	(Jia *et al*., 2017)
Orange	*CsLOB1* promoter	TG	Citrus canker	Citrus canker resistance	(Peng *et al*., 2017)
Wheat	*EDR1*	TG	PM	Powdery mildew resistance	(Y. Wang *et al*., 2014)
Wheat	*TaMLO*	TG	PM	Resistance to PM	(Y. Wang *et al*., 2014)
Wheat	*TaEDR1*	TG	PM	Resistance to PM	(Yunwei Zhang *et al*., 2017)
Rice	*OsSWEET11, OsSWEET14*	TG	Bacterial blight	Indels in promoter, resistance not confirmed	(Jiang *et a*l., 2013)
Rice	*OsMPK5*	TG	Fungal (*Magnaporthe grisea*) and bacterial (*Burkholderia glumae*) pathogens	Indels in promoter, resistance not confirmed	(Xie & Yang, 2013)

Tomato	*DMR6*	TG	*Pseudomonas syringae, Phytophthora capsici*, and *Xanthomonas spp*.	Broad spectrum resistance to specified bacterium	(Paula de Toledo Thomazella *et al*., 2016)

Bo2g016480 encodes a GDSL-Lipase, a family of intracellular membrane-bound proteins that has been found to play a role in intracellular signalling and regulation of plant defences. Examples of this include *AtGLIP1* in *Arabidopsis* which modulates resistance to *Alternaria brassicicola* in association with ethylene signalling of systemic acquired resistance (SAR) (Kwon *et al*., 2009). In rice (*Oryza sativa*), *OsGLIP1* and *OsGLIP2* were identified as functional lipases whose function in lipid metabolism was shown as a negative modulator of host immune responses to bacterial blight caused by *Xanthomonas oryzae pv. oryzae* and rice blast caused by *Magnaporthe oryzae* (Gao *et al*., 2017). As the obvious candidate dominant susceptibility factor, the A12DH Bo2g016480 allele provides an ideal target for a loss-of-function (knock-out) experiment to determine its role in host susceptibility to the race 9 *A. candida* isolate AcBoWells.

Interestingly, screening of ACA2 based markers on F_2 individuals from the A12DH x EBH527 mapping population revealed a rare recombinant line (EH177) that exhibits complete resistance to AcBoWells yet contains the A12DH haplotype across the ACA2 interval. This prompted prediction of a dominant major effect *R*-gene that is expected to be tightly linked to ACA2.

In the current study, investigation of the *B. oleracea* A12DH x EBH527 resistance is continued with two specific objectives: 1) CRISPR gene editing of the GDSL Lipase (Bo2g016480) to test whether loss-of-function mutation confers resistance to *A. candida* race 9; and 2) mapping of a dominant resistance locus using whole genome sequencing and bulked segregant analysis (WGS-BSA).

3.2 METHODS

3.2.1 Objective 1. CRISPR knock-out of candidate susceptibility gene Bo2g016480

3.2.2 Design of CRIPSPR sgRNA baits. Target sequences conforming to G(N)20GG for sgRNA bait design were identified on sense and antisense strands of the A12DHd Bo2g016480 allele first exon, using Geneious v10.2.3 (Kearse *et al*, 2012). These were then assessed for potential off-target matches within the *B. oleracea* TO1000DH3 reference assembly (Parkin *et al*., 2014) using BLAST searches (The *Brassica* Database. http://brassicadb.org/brad/blastPage.php). Final targets were chosen based on their off-target specificity as well as proximity to the start codon. Two tandem pairs of sgRNAs were chosen to target the first exon (Table 3.2) and thereby maximise the chances of generating a successful edit (Figure 3.1). To guaranty effective genotyping of the edits in downstream screening, suitable PCR primers were established prior to development of sgRNA constructs that cleanly amplify the edited region. These were the forward primer Ed_F (5-TTTCTTGGACCACAAATACTTTGGTAATAT-3) and the reverse primer Ed_R (5-TCAGAACAACTTGAGTTATTTCATTCTCAT-3).

Table 3.2. Designed sgRNAs, split between the two constructs 1991 + 1992 that each target the first exon of ACA2 candidate gene Bo2g016480. The sgRNAs are attached to the Cas9 DNA cleaving enzyme to induce functional knockouts of the target recessive GDSL-Lipase Bo2g016480. Black letters represent the spacer, which is complimentary to the target sequence. Red letters represent the protospacer adjacent motif (PAM) which is required for binding of the *Streptococcus pyogenes Cas9* endonuclease used here.

Construct	sgRNA-1	sgRNA-2
1991	GCGGTATCAACTAGAAGACCCGG	GGAGACTCGCTCGTCGACAGTGG
1992	GACCGGACACAACCAATGCTAGG	GGAAAATCAATTCCATATGGTGG

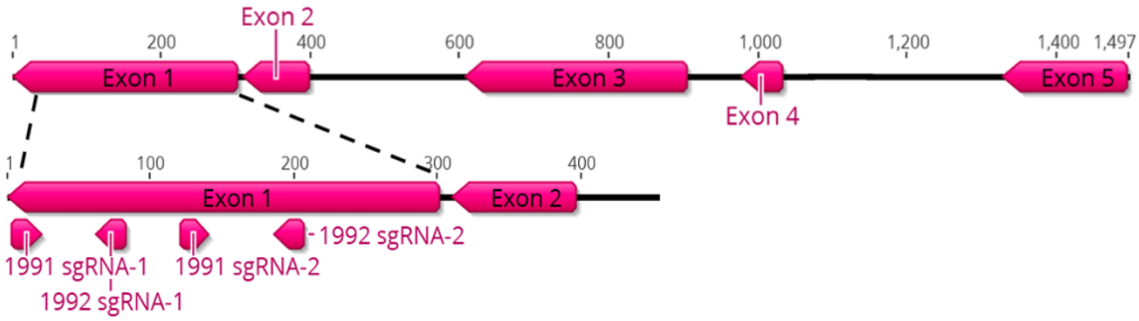


Figure 3.1. *Brassica oleracea* gene model for Bo2g016480, which includes five exons as determined by genome annotation of the T01000 reference assembly. The magnified lower portion displays the relative positions of sgRNA editing sites within exon 1, where 1991 and 1992 represent different binary constructs each containing two tandem sgRNAs (1 and 2 respectively).

3.2.3 Vector construction. CRISPR construct assembly was undertaken at the John Innes Centre in Norwich by the BRACT research facility. To compile the binary plasmid vector, Golden Gate Modular Cloning toolkit (Engler et al., 2008) and Addgene assembly components were assembled as described in (Lawrenson et al., 2015). The final pBRACT vector (sgRNABolC.GA4.a), containing 35 s tandem sgRNAs, a *Cas9* expression cassette and a spectinomycin resistance cassette can be visualised in Figure 3.2. The binary plasmid vector (sgRNABolC.GA4.a) is available online from the non-profit plasmid depository AddGene (https://www.addgene.org/browse/article/14759/).

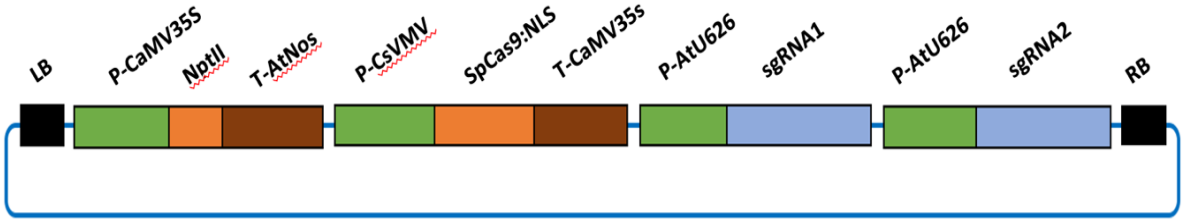


Figure 3.2. The binary plasmid vector sgRNABolC.GA4.a delivered to the transformable *Brassica oleracea* background DH1012. Transcription units were assembled into the binary plasmid backbone pAGM8031 using Golden Gate Modular Cloning. The vector houses a spectinomycin resistance cassette consisting of the *neomycin phosphotransferase* coding sequence (*nptII*) driven by a 35 s promoter P-CaMV35S and terminated by *T-AtNos*; the Cas9 component is driven by a constitutive promoter from Cassava Vein Mosaic Virus (P-CsVMV) and a tandem pair of sgRNAs (assembled into two cassettes - 1991; sgRNA1 + sgRNA2; 1992; sgRNA1 + sgRNA2 respectively) driven by an *Arabidopsis* U626 promoter (P-AtU626).

3.2.4 Brassica transformation. The transformable *B. oleracea* line DH1012 was used as a background for editing of the A12DH Bo2g016480 allele. This is a doubled haploid progeny from the *Brassica oleracea ssp alboglabra* (A12DHd) and *B. oleracea ssp italica* (GDDH33) mapping population (Bohuon *et al*, 1996) and is homozygous for the target allele. Transformation of DH1012 was performed by the BRACT research facility using *Agrobacterium tumefaciens* mediated infection of four-day-old cotyledonary petioles according to methods developed by (Hundleby & Irwin, 2014). The primary steps taken to maintain transformed juvenile plantlets can be seen in Figures 3.3 and 3.4, where explant transgenic shoots were isolated after three - four weeks of spectinomycin selection (15mg/l). Transgenic T_0 explants were screened for edits using DNA extraction and PCR amplification methods described by (Lawrenson *et al.*, 2015) and target specific primers Ed_F and Ed_R. Estimations of transgene insert copy number were determined for each explant using quantitative real-time PCR as described by (Weng *et al*., 2004).

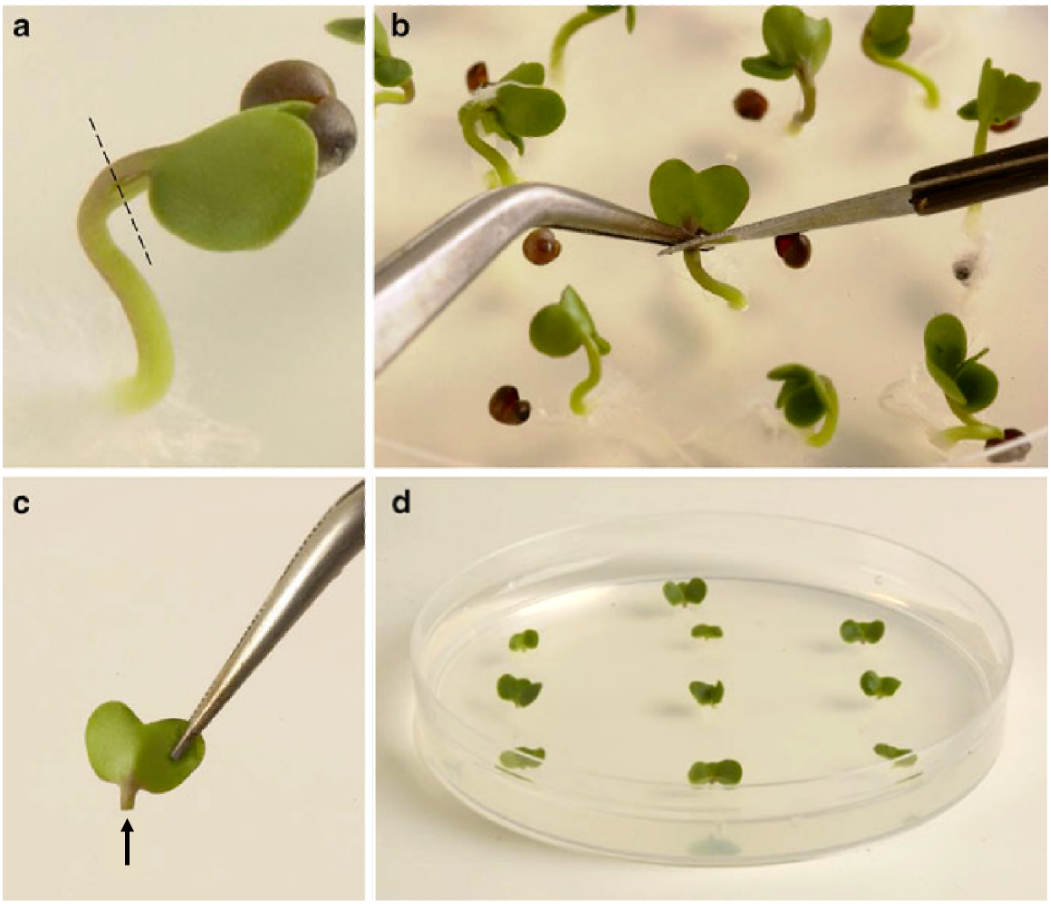


Figure 3.3. Explant isolation of four-day-old *Brassica oleracea* DH1012 seedlings from cotyledon petioles. **A)** The position where the petiole is exercised. **B)** Explant isolation. **C)** Cotyledonary petiole dipped into *Agrobacterium* suspension for transformation of the CRISPR/Cas9 t-DNA. **D)** Selection and maintenance of transformant explants on co-cultivation medium containing 15mg/l of selective antibiotic spectinomycin.

3.2.5 Processing transformants. A total of 96 T_0 transformant explants, established on spectinomycin selective growth media were then taken to Wellesbourne Crop Centre for continued development. Those showing elongated shoots were transferred to sterile peat plugs in magenta pots and kept in a Conviron growth cabinet at 20±2°C with a 10h photoperiod. Once established, these were transplanted to Levington M2 compost and maintained in shaded propagators for the first week to acclimatise to natural light intensity and decreased humidity. Adult plants producing buds were then sleeved with clear, perforated bags (Cryovac (UK) Ltd), and gently shaken daily to facilitate self-pollination (Figure 3.5). Pods were dried on the plants before being harvested and threshed for seed.

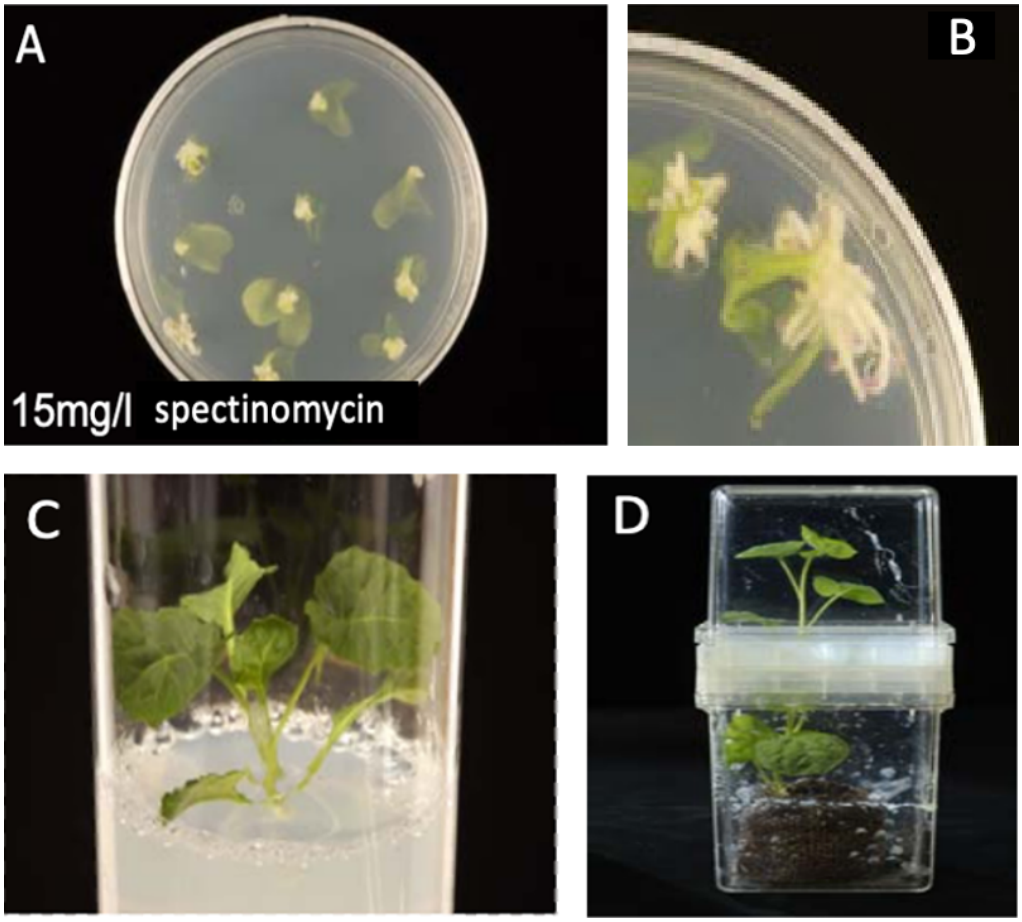


Figure 3.4. Shoot isolation of CRISPR transformed plantletes. A) Early formation of selection shoots after two weeks of maintenance on spectinomycin medium. B) Callus formation and evident chlorosis of non-transformed tissue under spectinomycin selection. C) Established T_0 juvenile plantlets, grown on Gamborgs B5 medium with increased (50mg/l) concentrations of spectinomycin to minimise development of escapee shoots that could result in chimeric plants. D) Plants transferred to sterile peat plugs in magenta pots, hanving shown sufficient root and shoot growth.

Figure 3.5. Self-fertilisation of CRISPR edited T_0 DH1012 lines in the glasshouse for production of T_1 seed with the following genotype characteristics: 1) a stably inherited homozygous edit; 2) segregated loss of the transgene cassette.

3.2.6 Molecular genotyping of transformants.

To genotype transformant seedlings, a single 6 mm leaf disk punch of tissue was harvested from primary leaves for DNA extraction using the CTAB method (Doyle, 1991). PCR amplification of an 820 bp product spanning the edited regions of Bo2g016480 was produced for all samples using the exon 1 edit-spanning primers Ed_F and Ed_R. Presence or absence of the transgene was established using *Cas9* specific primers - forward primer CasF (5-CTGCGAGTGAACACGGAGATC -3) and reverse primer CasR (5-AGAGAGTGTTTAGGAAGCACC-3). PCR reactions were conducted using Phusion high-fidelity DNA polymerase. Reaction volumes of 25 ul were produced using 5 µl of 5 x GC buffer, 0.5 µl of 10 mM dNTPs, 1.25 µl of 10 µm forward and reverse primers, 1 µl of template (~20 ng/ µl), 15.75 µl nuclease-free water and 0.25 µl of DNA polymerase added last. All reactions were prepared on ice. PCR was performed using the standard touchdown program (Table 3.2). A 3 µl sample of PCR product was added to 1 µl of loading dye for amplicon assessment using electrophoresis with a 1% agarose gel in Tris Borate buffer (TBE). Where required, PCR products were purified using the QIAquick PCR Purification Kit according to protocol.

A 5 ul aliquot of purified PCR product and 5 ul of primer at a concentration of 0.5mM was sent for Sanger sequencing through the LightRun service provided by GACT Biotech. Sequence analysis was performed using Geneious v10.2.3 (Kearse *et al.*, 2012). Trimmed reads were aligned to the DH1012 reference allele and visually inspected to detect polymorphisms.

Table 3.3. Standard PCR program used for the amplification of the Bo2g016480 CRISPR/Cas9 edited region, as well as the *Cas9* transgene. Where phases D = denaturation, A = annealing and E = extension.

	Start	(32 cycles)			Finish
	D	D	A	E	E
Temperature (°C)	98	98	60	72	72
Time (minutes)	5.00	0.30	0.30	1.30	7.00

3.2.7 Phenotypic assessment of confirmed edited lines. Thirty T_2 offspring from the advanced T_1 line 3.1, which possesses a homozygous edit and no transgene were assessed for altered susceptibility to AcBoWells. Unedited DH1012 was used as a susceptible control, and EH527 as a resistant control. All seedlings were tested using the same method previously described for the maintenance of *A. candida* isolates.

3.2.8 WGS-BSA, using segregating dominant resistance in recombinant EH177 (Objective 2). All methods applied in Objective 2, including production of EH177 phenotype bulks, WGS and BSA have been described already in Chapter 2 methods (Sections 2.2.3, 2.2.4, 2.2.6 and 2.2.7).

3.3 RESULTS

3.4 Objective 1. CRISPR knock-out of candidate susceptibility gene Bo2g016480

3.4.1 Management and selection of CRISPR edited transformants. In total, 96 T_0 transformant plants were provided by BRACT and taken for development at the University of Warwick Crop Centre. Genotype data provided with the plants was assessed to identify 22 individuals that contained distinct heterozygous edits. Of these, ten plants contained edits from 1991 – sgRNA 1, three from 1991 – sgRNA 2, six from 1992 – sgRNA 1, and three from 1992 -sgRNA 2. The remaining 74 plants contained transgenes though no edits were observed and the plants were subsequently discarded. Of the 22 lines containing active edits, six were previously described by quantitative real-time PCR as containing single transgene inserts (Weng *et al.*, 2004), four of which; 3.1, 8.2, 10.4 and 29.1 were advanced to the T_1 generation. Genotyping of the T_1 offspring (24 individuals per line) revealed that only line 3.1, showed a 3:1 segregation ratio for presence/absence of the transgene, whilst the others showed patterns consistent with a >1 copy number.

Of the T_1 offspring from line 3.1 that lacked the transgene, only two plants were found to contain homozygous edits, found at the cut site of 1992- sgRNA 1. In both cases, a single T insertion 66 nucleotides from the start codon of exon 1 generates a frame shift mutation. This introduces a premature stop codon nine amino acids downstream from the edit, as seen in the ExPASy protein translation (Figure 3.8), resulting in a substantially truncated version of the DH1012 Bo2g016480 protein.

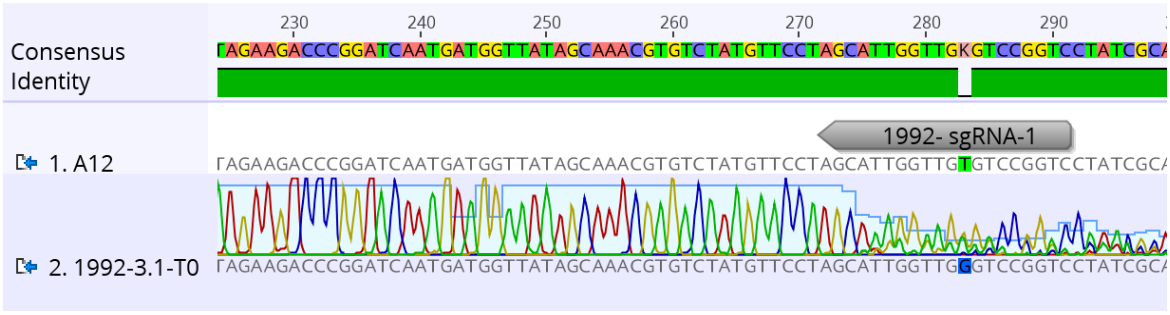


Figure 3.6. Sanger chromatogram showing the position of the initial heterozygous edit made by sgRNA-1 (construct 1992) in the first exon of T_0 transformant line 3.1. This is evident by the 3' introduction of mixed sequence template.

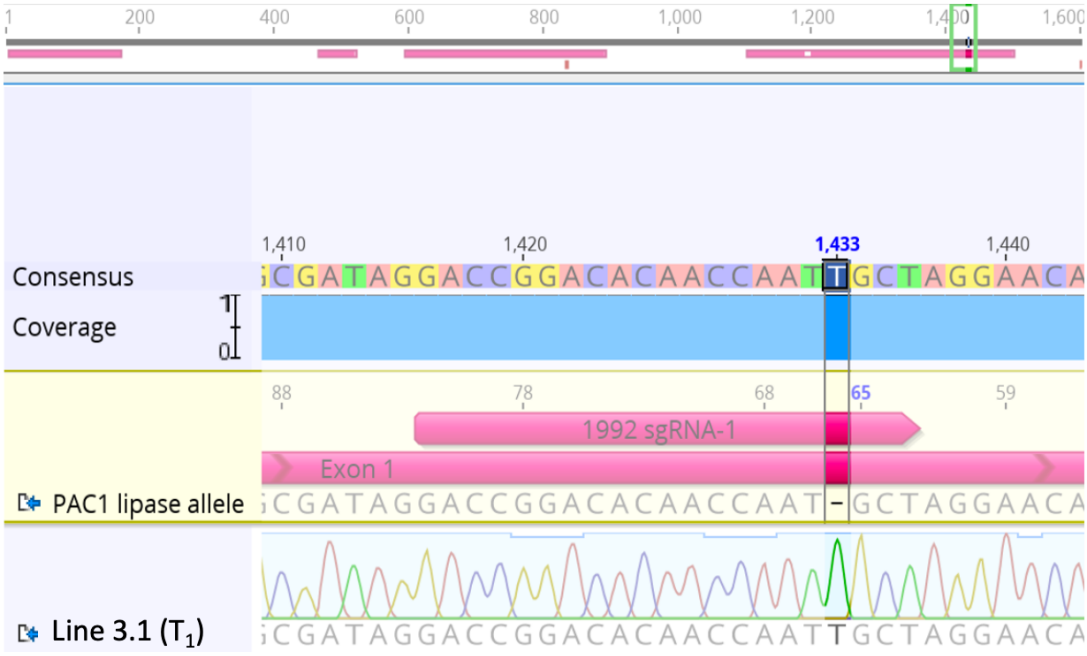


Figure 3.7. Chromatogram showing stable inheritance of a single homozygous T insertion (position 1433) in the transgene free T_1 line 3.1. The self-fertilised T_2 offspring from this line then contain the fixed edited genotype required for the Bo2g016480 loss-of-function experiment.

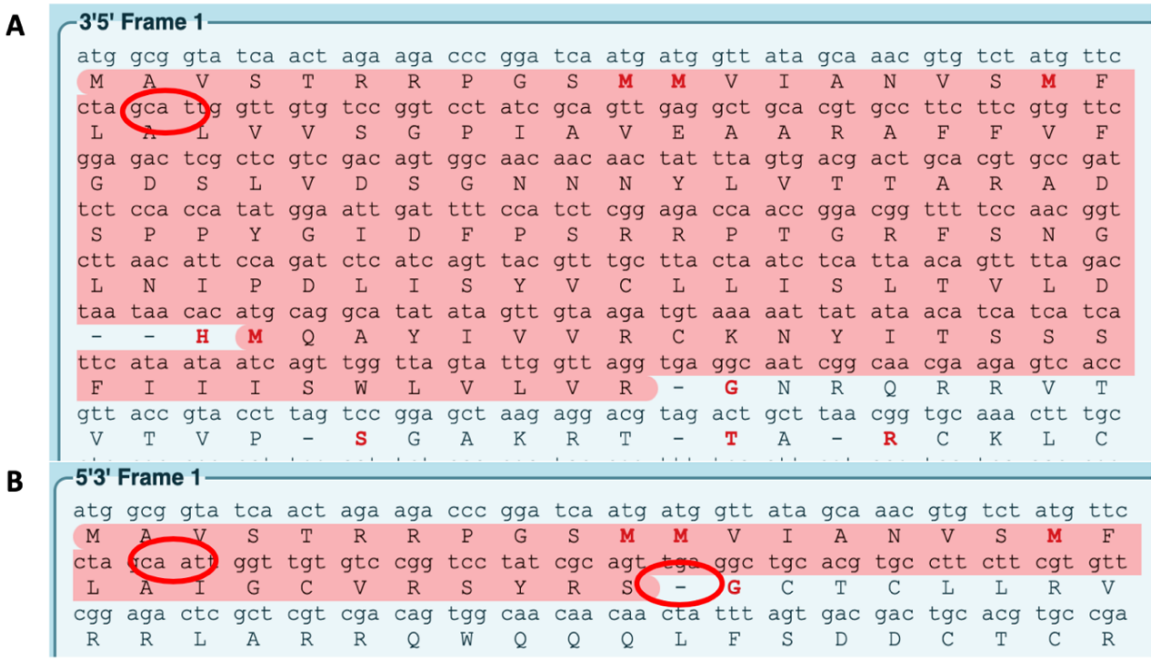


Figure 3.8. Translation of the DH1012 Bo2g016480 first exon to amino acid sequence. **A)** Complete first and second exons of the unedited DH1012 (A12DH) allele. **B)** Translation of the first exon for edited line 3.1, where the single T (reverse transcribed here) insertion introduces a frame shift mutation which results in a downstream premature stop codon.

3.4.2 Phenotypic assessment of CRISPR edited line 3.1 for altered susceptibility to race 9 isolate AcBoWells.

If Bo2g016480 is a functional susceptibility gene in DH1012, then a homozygous knock-out mutation of the DH1012 allele is expected to confer resistance to *A. candida* race 9 isolates. A total of 30 T_2 offspring from the homozygous edited (and transgene free) line 3.1 were assessed for reduced susceptibility to isolate AcBoWells, relative to control plants of wild-type DH1012. The results in Figure 3.9 show that the knockout of the Bo2g016480 allele caused no difference to the susceptibility phenotype. Tested plants were genotyped and confirmed to contain the homozygous knockout, as well as no transgene. All mutant seedlings were morphologically normal compared to wild-type, indicating that early development of DH1012 plants was not affected by the Bo2g016480 knockout.

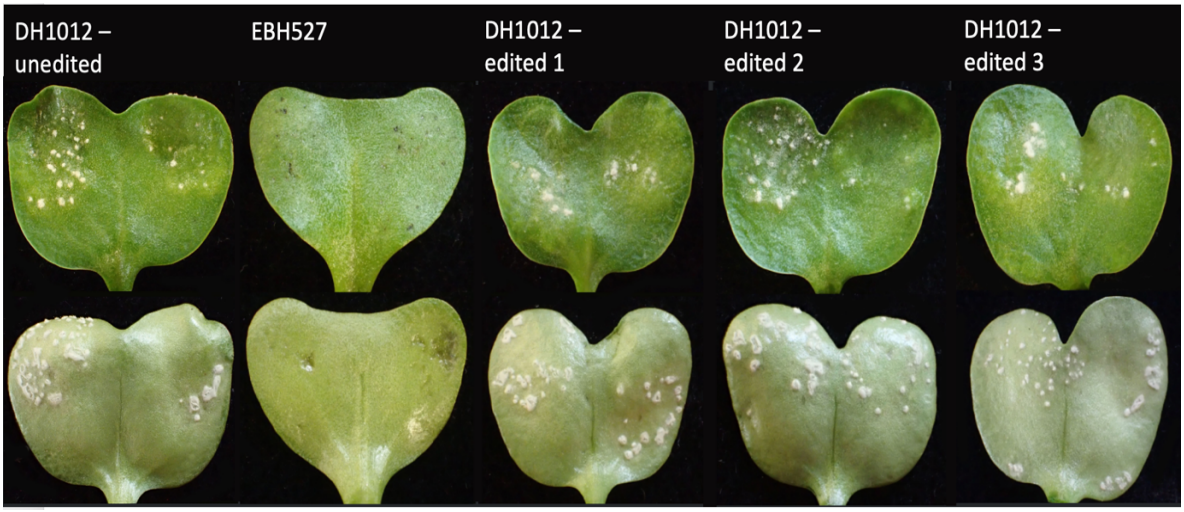


Figure 3.9. Interaction phenotypes observed on edited DH1012 line 3.1 that contains a homozygous knock-out of the candidate Bo2g016480 allele, when infected (ten days post inoculation) with the race 9 isolate AcBoWells (ten days post inoculation). This is shown relative to the unedited DH1012 positive control and the EBH527 negative control. No alteration to susceptibility is found in edited plants, suggesting that Bo2g016480 is not the ACA2 recessive susceptibility gene.

3.4.3 WGS of parents EBH527 and A12DH for complete sequence capture of the ACA2 locus.

To search for other possible causal mutations within the ACA2 18 kb locus, DNA from the resistant EBH527 and susceptible A12DHd parents were sent for WGS. Illumina sequencing was performed at 30x sequencing coverage, generating a total of 19.9 Gb of data for EBH527 and 18 Gb for A12DH. This equated to approximately 132 and 120 million paired-end reads respectively, which were quality assessed in FastQC (Andrews, 2010) prior to alignment to the *B. oleracea* TO1000DH3 reference assembly (Parkin *et al.*, 2014). The outputted *.bam files were visualised in IGV (v2.5.3) across the chromosome 2 ACA2 interval and all mutational differences between the two parents are reported in Table 3.4. Sequence coverage across the ACA2 interval was approximately 90% complete, with full coverage of the four ACA2 genes.

Table 3.4. Complete set of markers across the fine mapped 18 kb ACA2 interval, generated from WGS datasets of the *Brassica oleracea* resistant (EBH527) and susceptible (A12DH) parents. Marker ID includes the chromosome number (C2) and physical position (bp). CDS, indicates weather the marker is within coding sequence of a gene, as opposed to intron or intergenic sequence.

Marker ID	Type	CDS?	Gene ID	EBH527	A12DH	Description
C2_4846936	SNP	Y	Bo2g016480	A	T	Exon no. 5
C2_4854033	INDEL	N	-	TA insert	T	106 bp 5' of Bo2g016490
C2_4854092	SNP	N	-	A	G	165 bp 5' of Bo2g016490
C2_4855140	SNP	N	-	C	T	1215 bp 5' of Bo2g016490
C2_4855544	SNP	N	-	C	A	1619 bp 5' of Bo2g016490
C2_4856205	SNP	N	-	T	C	2280 bp 5' of Bo2g016490
C2_4858031	SNP	N	-	G	C	1436 bp 5' of Bo2g016500
C2_4858972	SNP	N	-	T	C	496 bp 5' of Bo2g016500
C2_4859422	SNP	N	-	C	T	45 bp 5' of Bo2g016500
C2_4863827	SNP	Y	Bo2g016510	C	T	Within single intron
C2_4863839	SNP	Y	Bo2g016510	C	T	Within single intron
C2_4864181	SNP	Y	Bo2g016510	A	C	Within single intron

The ACA2 interval contains three other possible candidate genes; Bo2g016490 (an ethylene-responsive transcription factor), Bo2g016500 (a protein of unknown function)

and Bo2g016510 (a G-type lectin S-receptor-like serine/threonine-protein kinase). Only Bo2g016510 (which was previously reported to be monomorphic) was found to contain mutational differences between the parents. However, the three SNPs within CDS all result in synonymous bp changes with no predicted alteration to amino acid sequence of the polypeptide. Both Bo2g016490 and Bo2g016500 are confirmed here as monomorphic, though each gene contains mutations within 5' UTR sequences. Within 200 bp of the start codon of Bo2g016490, EBH527 contains a 2 bp TA insertion at position 4854033 and a SNP at position 4854092. Whereas Bo2g016500 has two SNPs, 46 bp and 496 bp 5' of the start codon, at positions 4859422 and 4858972, respectively.

3.4.4 Development of ACA2 markers, using recombinant inbred line RIL_59 to assess remaining candidate genes. To further narrow the 18kb ACA2 interval and potentially rule out any of the other three remaining candidate genes (Bo2g016490, Bo2g016500, Bo2g016510), it is necessary to screen internal ACA2 markers (Table 3.4) using individual lines that are still segregating within the locus. Previous analysis by Dr Fairhead identified the resistant F_5 RIL accession RIL_59 (A12DH x 527EBH) as one such recombinant using flanking markers that originally determined the physical mapped ACA2 interval. Here, three pairs of primers were designed from the parent WGS datasets, to amplify internal ACA2 markers across the interval that distinguish each of the three candidate genes (see Table 3.5). This was undertaken using accession RIL_59, as well as the two parents.

Table 3.5. Primers for amplifying markers developed within the ACA2 mapping interval, used to genotype the three-remaining candidate ACA2 genes in recombinant accession RIL_59 as well as parents A12DH and EBH527.

Amplicon ID	Primers	Markers amplified	Closest gene	Clean amplification?
Amp_490	5-CGTTGCCTAACTCCACGGTA-3 5-GTTTGCCAGGGCATCCTCTA-3	C2_4854033 C2_4854092	Bo2g016490 (Intragenic)	Yes
Amp_500	5-TCTGTCGTCGCCATTACTGT-3 5-TCTCAAGCTGCTGACTTGTC-3	C2_4858972 C2_4859422	Bo2g016500 (Intragenic)	Yes
Amp_510	5-GTTGATCTCCTCGGGTGG-3 5-TCAGTACTCCGCTTGACCAG-3	C2_4863827 C2_4863839 C2_4864181	Bo2g016510 (CDS)	Only in A12DH

Genotype results from this experiment were unable to rule out any of the three remaining candidates however, as RIL_59 was found to contain a resistant (EBH527)

haplotype across the markers that distinguish Bo2g016490 (Amp_490) and Bo2g016500 (Amp_500) (Figure 3.10). Whilst the marker inside of Bo2g016510 (Amp_510) failed to cleanly amplify and so this candidate cannot yet be ruled out either. Ultimately, with no other ACA2 recombinants available, independent CRISPR knock-outs of the three candidate genes would be the next requirement for identifying the causal gene.

Accession	Phenotype	C2_4567864	Bo2g016470	Bo2g016480	Amp_490	Amp_500	Amp_510	Bo2g016520	Bo2g016550	C2_4978392
EBH527	R	AA	AA	AA	AA	AA	AA	AA	AA	AA
A12	S	BB	BB	BB	BB	BB	BB	BB	BB	BB
RIL_59	R	AA	AA	AA	AA	AA	?	BB	BB	BB
RIL_49	R	AB	AB	AA	AA	AA	AA	AA	AA	AA
RIL_38	S	AB	BB	BB	BB	BB	BB	BB	BB	BB
RIL_22	S	AA	AB	AB	AB	AB	AB	AB	AB	AB

Figure 3.10. Summary of interaction phenotype and genotype information for key recombinants from previous mapping of the ACA2 locus with recombinant inbred lines (RIL) on chromosome 2 of *Brassica oleracea* (Fairhead, 2016). The F_5 RIL population was produced from a cross of A12DH (susceptible) and EBH527 (resistant). New genotype data is shown within the black outlined block. Internal ACA2 markers were identified from WGS of the two parents and three PCR amplicons (Amp_490, Amp_500 and Amp_510) were produced that span several of these markers. Marker positions relative to the three remaining candidate genes (Bo2g016490, Bo2g016500 and Bo2g016510) can be inferred from Table 3.4. These were screened on RIL_59, which represents the only RIL available that recombines inside this region where haplotype data could possibly rule out remaining candidates. The remaining RILs documented here (RIL_49, 38 and 22) were genotyped in the previous study with ACA2 flanking markers and have been included to clarify genotype x phenotype relationships across the window. Results show however that neither Bo2g016490 or Bo2g016500 can be ruled out as RIL_59 contains the EH527 haplotype across this region. Bo2g016510 cannot be ruled out because marker amplification failed to produce quality sequence data for assessment of genotype. Phenotypes were determined by inoculation of inbreds with a race 9 isolate of *Albugo candida* (AcBoWells) and assessment after ten days. For genotype scores: AA indicates homozygous for an EBH527 allele; BB indicates homozygous for an A12DH allele; and AB indicates heterozygous alleles.

3.5 Objective 2. Mapping a dominant white rust resistance in *Brassica oleracea* using WGS-BSA.

3.5.1 Inheritance of white rust resistance in *Brassica oleracea* line EH177. The F_2 recombinant line EH177 (A12DH x EBH527) is predicted to contain a dominant resistance gene that is tightly linked to ACA2. Inheritance of this resistance was assessed in self-fertilised F_3 and F_4 generations. Segregation for resistance in the F_3 was found in a ratio of 289 resistant to 161 susceptible individuals, confirming dominance of the trait as well as heterozygosity of the F_2 parent plant. Two distinct phenotypes were observed: a resistant class showing minor necrotic cell death with no sporulation, and a susceptible class exhibiting similar levels of sporulation to the susceptible A12DH parent (Figure 3.11). No intermediate phenotype was observed. Chi-squared analysis ($\chi 2$) of the F_3 segregation ratio was tested for multiple gene models and an expected 2:1 ratio of resistant to susceptible provided the best fit for the observed data ($\chi 2 = 1.21$, $p = 0.3 < p < 0.2$), suggesting the presence of a lethal allele where homozygous recessive individuals do not survive embryonic development. However, phenotypic assessment of F_3 seedlings was made during winter in a polytunnel, which had a significant impact on seedling survival and the ability to decern clear symptoms.

To produce tissue bulks from distinct phenotypes the 50 F_3 lines were progeny tested to identify lines that were either fixed or still segregating for resistance, with 12 individuals sown and assessed per line. Of these, 13 lines were uniformly resistant and nine were uniformly susceptible, indicating homozygosity for either the resistant or susceptibility allele, respectively. The remaining 28 lines were still segregating (heterozygous). If considered in the context of a dominant single gene model (with 41 lines AA or AB, vs 9 lines BB), this generates an $\chi 2$ value of 1.3, and $p = 0.1 < p < 0.3$, with observed data meeting the expectations for this model and suggesting action of a single dominant gene.

Figure 3.11. Interaction phenotypes observed on the upper (top row) and lower (bottom row) cotyledon surfaces of segregating *Brassica oleracea* EH177 individuals following inoculation with a race 9 isolate of AcBoWells. Susceptibility (left) is comparable to that seen in the A12DH parent, and resistance (right) presents minor necrosis on the upper leaf surface.

3.5.2 Referenced-based SNP identification from Whole Genome Sequencing of tissue bulks.

Whole genome sequences (WGS) of pooled DNAs were generated using paired end (150 bp) sequencing on an Illumina Novaseq6000 platform by Novogene UK. Sequencing of the EH177 phenotype bulks (90x sequencing coverage) generated 120.9 and 124.5 M PE reads for the resistant and susceptible samples respectively. Low quality reads and adapters were removed before checking read quality using FASTQC (Andrews, 2010). High quality sequences were aligned and mapped to the *B. oleracea* TO1000DH3 reference assembly using BWA (v0.7.12, (Li & Durbin, 2009)) with default parameters. The sequence reads of the tissue bulks provided a combined average 57.3x coverage of the *B. oleracea* genome, with an average of 87.8% breadth of the reference genome covered by mapped reads. The average alignment rate of reads to the reference assembly was 85.15%, with an average of 77.12% mapping in pairs.

The GATK (Genome Analysis Toolkit) pipeline (Figure 2.1) was followed for reference-based SNP identification, with the application of HaplotypeCaller for calling SNPs in each bulk (Poplin *et al.*, 2017). These were then merged for the resistant and

susceptible sample, with a combined total of 3649273 SNPs produced across the nine *B. oleracea* chromosomes.

3.5.3 WGS based BSA mapping of a major locus for white rust resistance in *B. oleracea* line EH177.

For QTL mapping of loci affecting EH177 resistance, WGS-BSA was performed using QTLseqr (Mansfeld & Grumet, 2018), and the G' approach (Magwene *et al*., 2011). Firstly, consolidated resistant and susceptible SNPs were passed into QTLseqr for continued filtering of low confidence markers. This was performed where read depth was < 40, where reference allele frequency was < 0.3 (removing SNPs that are over- or under-represented in both bulks), and where a GATK GQ score was < 99. Reads with large discrepancies between the bulks were removed (> 100 depth difference), which could otherwise affect calculation of the G statistic. From the original SNP number of 3649273, 3391638 were removed to leave 257635. Filtering thresholds were established by plotting histograms based on SNP quality parameters (Figure 3.12).

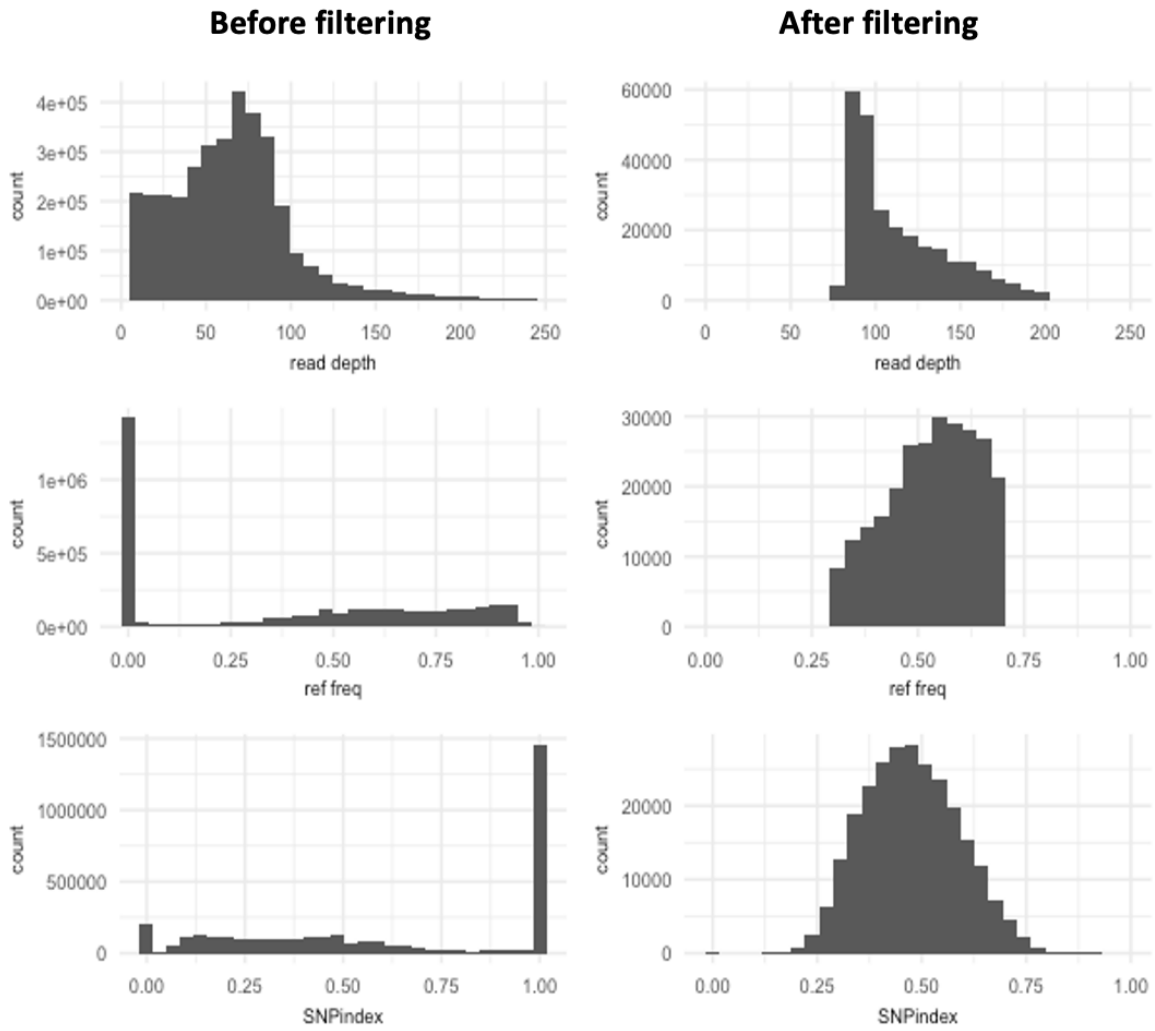


Figure 3.12. Quality filtering of WGS SNPs, based histogram distributions of raw read metrics that include: A) total read depth, before and after filtering reads < 40; B) Total reference allele frequency, before and after filtering reads < 0.3; C) SNP-index data before and after applying filtering steps. High-quality data should be approximately distributed around 0.5 in an F_2 population, as observed post-filtering.

A SNP index and G statistic was calculated for each individual SNP, as described by (Magwene *et al.*, 2011). The tricube smoothed delta SNP index and G value (G' value) were then calculated within a window size spanning 1.0 Mb of genomic region and were plotted against all nine *B. oleracea* chromosomes (Figure 3.14). The G' statistic functions as a weighted moving average across neighbouring SNPs, accounting for linkage disequilibrium (LD), whilst also minimising noise attributed to SNP calling errors (Mansfeld & Grumet, 2018). G' values calculated for each SNP are weighted by physical distance to the focal SNP, decreasing in value as they get closer to the window edge. Significance thresholds (*p*-values), and genome-wide Benjamini-Hochberg false discovery rate (FDR)

(Hochberg, 2016) adjusted *p*-values (*q*-values) were estimated from the null distribution of the G', which assumes there is no QTL linked to the SNP (Figure 3.13). The tricube smoothed G' values from the analysis of EH177 bulked resistance showed significant G' peaks on chromosomes C2 and C6, with peaks above the FDR (q) of 0.001, suggesting these regions most likely contain the loci for WRR in line EH177 (Figure 3.14). These QTLs will be referred to as Bol_EH177_C2 (chromosome 2 QTL) and Bol_EH177_C6 (chromosome 6 QTL) respectively.

The negative direction of the delta SNP value for the Bol_EH177_C2 peak suggests that the association originates from the resistant bulk, whereas the Bol_EH177_C6 peak is ambiguous regarding which sample resistance originates from. The genomic region covered by the Bol_EH177_C2 peak is approximately 3.14 Mb, whilst the peak of Bol_EH177_C6 covers 3.15 Mb (Table 3.6). As predicted, the Bol_EH177_C2 peak is tightly linked to the ACA2 recessive locus, where the 3' interval edge is only ~66 kb from the candidate Bo2g016480 gene.

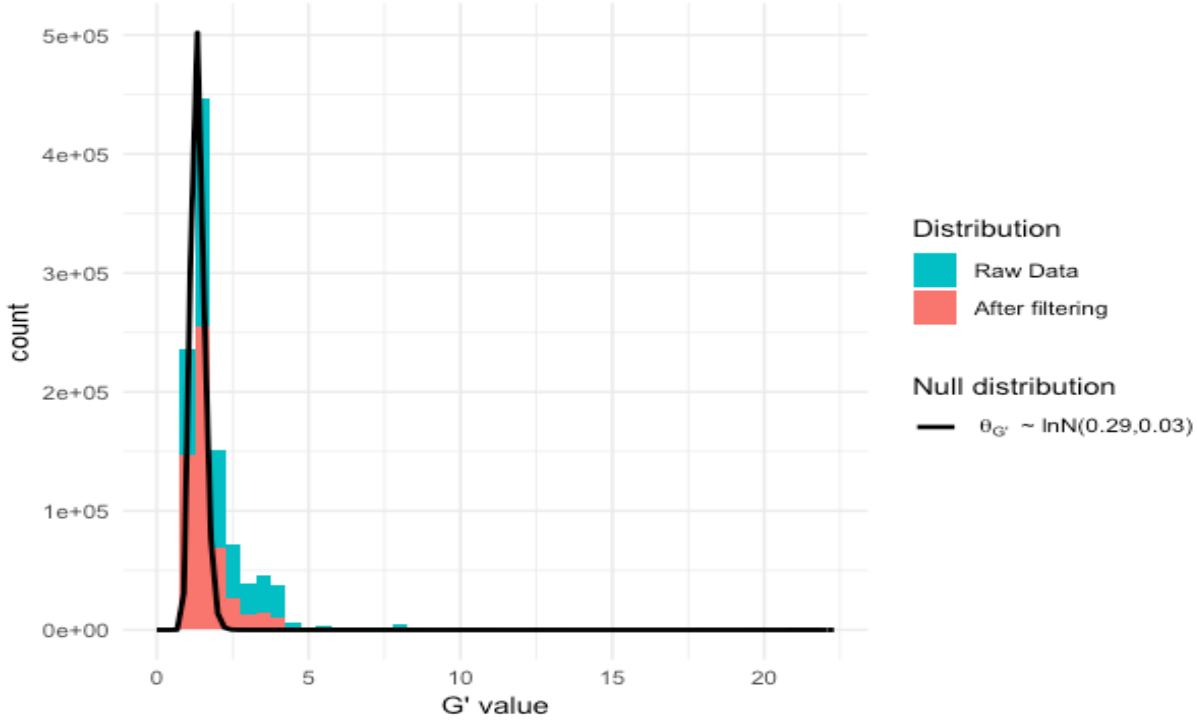


Figure 3.13. Tricube smoothed G' statistics, indicating a closer fit to the null distribution once low confidence SNPs have been removed. These might otherwise have adversely effected calculation of *p-values* which are estimated from the null distribution of G'.

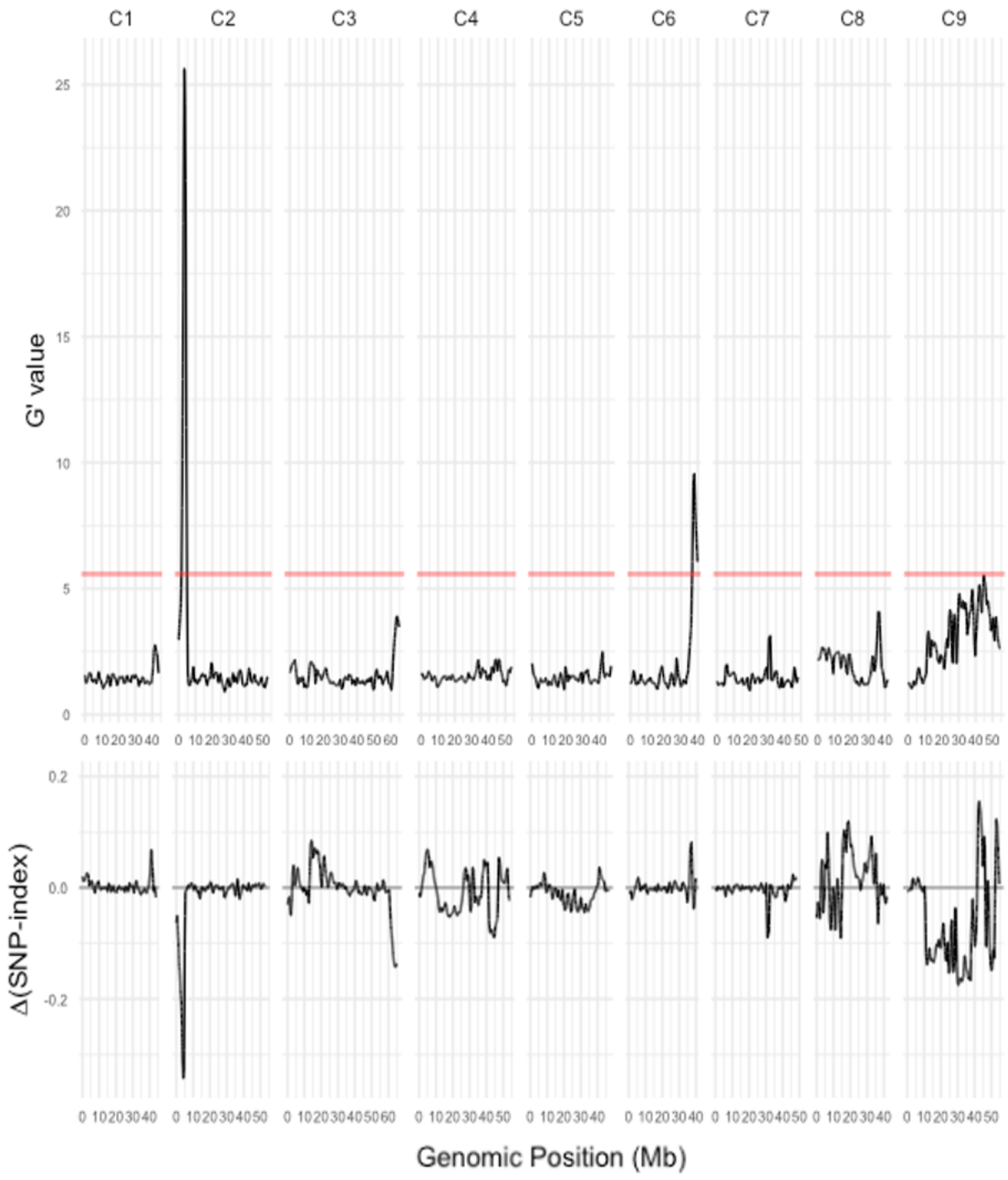


Figure 3.14. Quantitative trait loci (QTL) for white blister rust resistance in *Brassica oleracea*, shown by a G' value plot. Peaks are predominantly on chromosomes C2 and C6, identified in resistant line EH177 (A12DH x EBH527). Distribution of the G' value calculated with a 1-Mb sliding window using tricube smoothing kernel across the nine *B. oleracea* chromosomes based on the TO1000 genome assembly. The red line denotes the significance threshold FDR = 0.001. Genomic regions where the G' crosses the threshold value are considered as significant. Negative values in the delta SNP-index suggests that the allele contributing to the C2 (Bol_EH177_C2) peak is from the resistant bulk, whilst C6 (Bol_EH177_C6) is ambiguous.

Table 3.6. Summary of the two QTL peaks (FDR = 0.001) on C2 and C6 from the G' analysis of EH177 resistance. QTL – ID. Chr – chromosome. Start/End the bp position that passes the FDR threshold. Length –physical distance of the genomic interval (bp). peakDeltaSNP – the delta prime value at the peak summit. maxGprime – G' value at the peak summit. posMaxGprime - the genomic position of the maximum G' position in the QTL. meanPval – the average *p*-value in the region. meanQval – the average adjusted *p*-value in the region.

QTL	Chr	Start	End	Length	peakDeltaSNP	maxGprime	posMaxGprime	meanPval	meanQval
Bol_EH177_C2	C2	1633149	4781105	3.14E+06	-0.215	25.624	3310173	8.76E-06	5.33E-04
Bol_EH177_C6	C6	36666009	39817658	3.15E+06	-0.042	9.546	37952739	1.46E-05	9.70E-04

3.5.4 Candidate gene analysis of QTLs. Gene annotation files for the *B. oleracea* TO1000 reference show that the 3.1 Mb C2 interval Bol_EH177_C2 contains 771 genes in total. To identify candidate *R*-genes, Disease Resistance Analysis and Gene Orthology (DRAGO 2) software was applied to annotate LRR, Kinase, NBS and TIR domains in the reference (Osuna-Cruz *et al.*, 2018). Here, we identified a cluster of 13 resistance associated genes within the EH177 C2 interval, comprised of ten NLR genes and three receptor-like kinases (RLKs). When checking WGS across these genes, all except Bo2g011890 (monomorphic) were found to contain SNP differences within exons between the resistant and susceptible bulks. The TNL genes Bo2g014340 and Bo2g016440 however lack coverage and therefore cannot be confirmed as polymorphic (Table 3.7). Whilst the C6 QTL Bol_EH177_C6 contains nine conventional resistance gene candidates; including two TNL genes, five RLKs and two coiled-coil kinases (CKs) (Table 3.8).

Table 3.7. Resistance genes (*R*-genes) within with the EH177 chromosome 2 QTL Bol_EH177_C2, as well as those close to the 3' QTL edge proximate to ACA2. These include the TNL genes (Bo2g016440 + Bo2g016470) which contain resistance-associated SNPs (raSNPs) and could under different BSA mapping parameters be included within the QTL. No raSNPs are found downstream of Bo2g016470 (only the homozygous A12DH haplotype) suggesting this is the approximate cross-over position between A12DH x EBH527 in line EH177. In the raSNP column MS refers to missing sequence.

Gene ID	In QTL?	Type	Start Pos.	Length	raSNP?
Bo2g010410	Y	TNL	2461583	3628	Y
Bo2g010720	Y	TNL	2631754	4028	Y
Bo2g011890	Y	RLK	3300854	1917	N
Bo2g013170	Y	RLK	4027039	2132	Y
Bo2g013340	Y	RLK	4120171	6262	Y
Bo2g013810	Y	TNL	4380651	5419	Y
Bo2g013830	Y	TNL	4389266	4134	Y
Bo2g014060	Y	TNL	4547657	5101	Y
Bo2g014070	Y	NL	4556225	4174	Y
Bo2g014110	Y	TNL	4576614	26454	Y
Bo2g014320	Y	TNL	4717537	9556	Y
Bo2g014340	Y	TNL	4739657	4070	MS
Bo2g014350	Y	TNL	4745390	4243	Y
Bo2g016440	N	TNL	4812074	4778	Y
Bo2g016470	N	TNL	4830258	12313	Y
Bo2g016480	N	LIP	4846722	1496	N

Bo2g016610	N	CNL	4911474	3649	N
Bo2g0EH17760	N	TNL	4990764	2502	N

Table 3.8. *R*-genes within with the EH177 chromosome 6 QTL Bol_EH177_C6.

Gene ID	Type	Start Pos.	Length	raSNP?
Bo6g118660	RLK	37009978	3251	N
Bo6g118790	RLK	37107159	4126	Y
Bo6g119590	RLK	37475024	3542	Y
Bo6g119760	RLK	37610630	3425	Y
Bo6g119900	RLK	37685302	3035	Y
Bo6g122440	CK	39098444	3360	Y
Bo6g123720	TNL	39271601	3403	Y
Bo6g123970	TNL	39416187	3787	Y
Bo6g124030	CK	39438397	3887	N

3.5.5 Assessment of the EH177 C2 QTL relative to the recessive ACA2 locus.

The discovery of EH177 resistance in a rare F_2 recombinant, derived from the same mapping population (A12DH x EBH527) used to identify the recessive ACA2 locus, lead to the hypothesis that the two resistance loci were tightly linked. WGS-BSA mapping of the EH177 resistance locus confirmed this to be the case, identifying a highly significant QTL on chromosome 2, designated Bol_EH177_C2. The previously assessed recessive resistance candidate Bo2g016480 lies approximately 65 kb from the closest EH177 QTL boundary. To identify the approximate physical recombination position whereby EH177 resistance segregated from ACA2 resistance, GBS markers derived from the parents A12DH and EBH527 were assessed in IGV (v2.5.3) across the combined interval. This was located to a 2.79 kb region between markers at positions 4567864 and 4846936, where if moving 3' to 5', allele frequencies switch from homozygous for the A12DH parent to homozygous for the EBH527 parent, before resuming a state of increasing heterozygosity. A visual assessment of WGS SNPs was conducted in IGV to locate the approximate position between these two markers. A ~5 kb region, between the recessive GDSL-lipase ACA2 candidate Bo2g016480 and its 5' neighbour Bo2g016470 was identified as the site of recombination (see Figure 3.15).

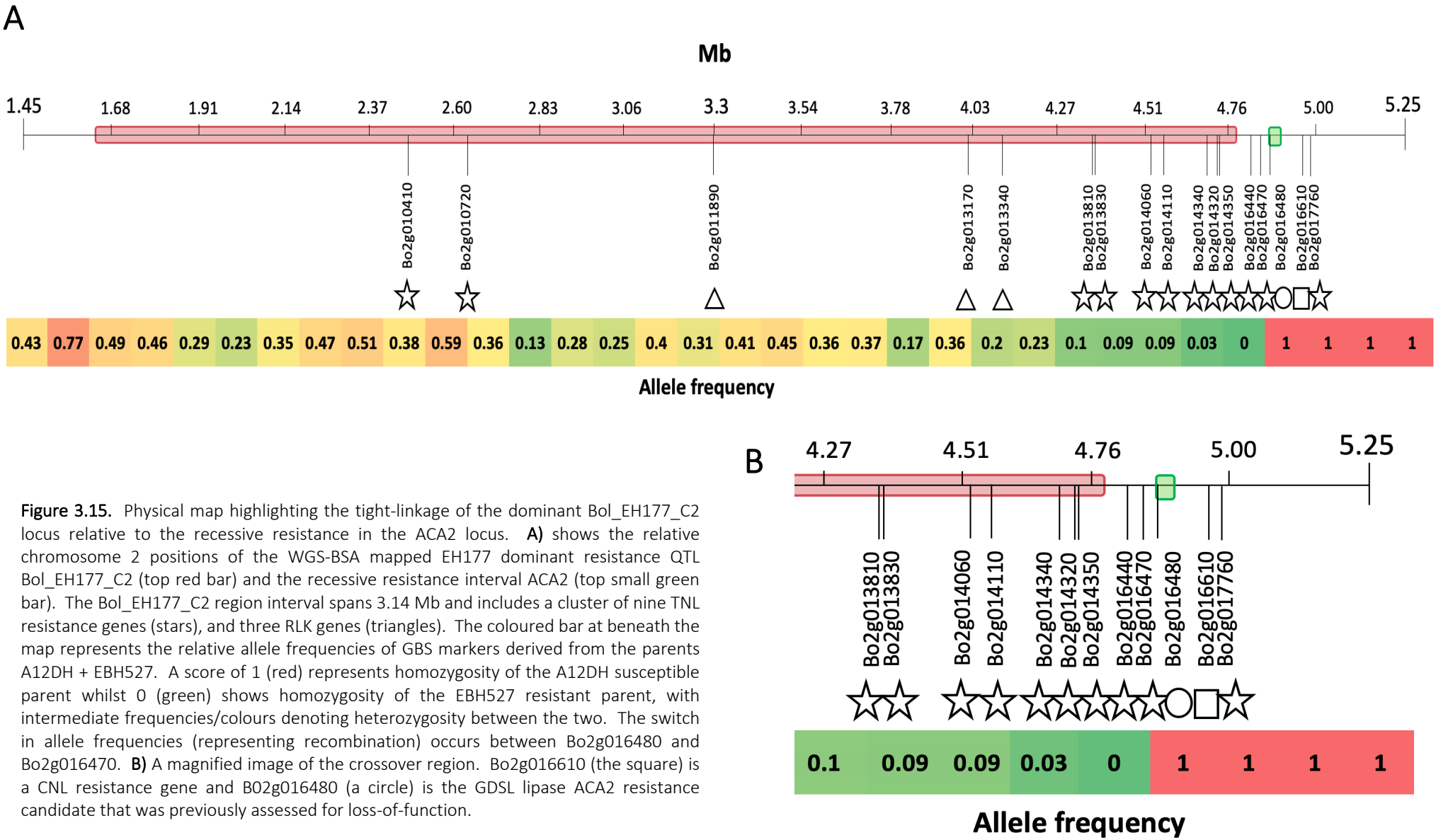


Figure 3.15. Physical map highlighting the tight-linkage of the dominant Bol_EH177_C2 locus relative to the recessive resistance in the ACA2 locus. **A)** shows the relative chromosome 2 positions of the WGS-BSA mapped EH177 dominant resistance QTL Bol_EH177_C2 (top red bar) and the recessive resistance interval ACA2 (top small green bar). The Bol_EH177_C2 region interval spans 3.14 Mb and includes a cluster of nine TNL resistance genes (stars), and three RLK genes (triangles). The coloured bar at beneath the map represents the relative allele frequencies of GBS markers derived from the parents A12DH + EBH527. A score of 1 (red) represents homozygosity of the A12DH susceptible parent whilst 0 (green) shows homozygosity of the EBH527 resistant parent, with intermediate frequencies/colours denoting heterozygosity between the two. The switch in allele frequencies (representing recombination) occurs between Bo2g016480 and Bo2g016470. **B)** A magnified image of the crossover region. Bo2g016610 (the square) is a CNL resistance gene and B02g016480 (a circle) is the GDSL lipase ACA2 resistance candidate that was previously assessed for loss-of-function.

3.6 DISCUSSION

3.6.1 Objective 1 discussion. Previous mapping of broad-spectrum white rust resistance to *A. candida* race 9 at the ACA2 locus predicted that a gene encoding a GDSL-lipase (Bo2g016480) was the functional determinant (Fairhead, 2016). The recessive nature of the ACA2 resistance implies that such a gene acts as a susceptibility factor, where both copies of the A12DHd allele are required by the pathogen to enable infection. To confirm the role of the candidate GDSL lipase in ACA2 resistance it was necessary therefore to knock-out function of the candidate susceptibility allele, with the expectation it would induce a EBH527 resistant phenotype. The CRISPR/Cas9 Bo2g016480 knockout produced here contains a frame shift mutation in the first exon which is predicted to induce a heavily truncated and non-functional GDSL-lipase allele. However, when tested, no alteration to the susceptibility phenotype was observed between the Bo2g016480 mutants and the unedited DH1012 background (which contains the A12DH Bo2g016480 allele), suggesting that Bo2g016480 is either not the causal gene or that its function is somehow dependent on genetic background (e.g., heterozygosity of another gene).

This was an unexpected result at the time in the light of previously established genetic evidence suggesting that the three other ACA2 candidates (Bo2g016490, Bo2g016500 and Bo2g016510) appeared to be monomorphic. To advance the investigation of the ACA2 locus, it was deemed necessary to obtain complete sequence across the mapped interval for both A12DHd and EBH527 parents, to detail all available mutations in both coding and non-coding regions. The acquired datasets revealed a number of interesting mutations, though ultimately ones that when tested for co-segregation with phenotype with the ACA2 recombining line RIL_59 cannot rule out any of the three remaining genes as potential candidates. Here, each remaining candidate will be considered based upon what is currently known about their function within plants, as well as their established mutational evidence.

Candidate gene Bo2g016490 is an ethylene responsive transcription factor (ERF), belonging to a family of important stress response regulators in plants that act as a key regulatory hub for biotic and abiotic stress responses (Müller & Munné-Bosch, 2015). ERFs represent one of the largest subfamilies of plant transcription factors that are characterised by a single, highly conserved AP2 domain. They act by binding to

downstream regulatory elements that include cis-acting AGCCGCC motifs (the GCC box) to modulate expression of other stress response genes. ERFs have been implicated as a 'double-edged sword' in plant defences, in some instances activating stress responses whilst in others acting as repressors (Thirugnanasambantham *et al*., 2015). For example, in *Arabidopsis* the ERF5 protein is able to increase resistance to the bacterial pathogen *Pseudomonas syringae pv. tomato* via salicylic acid signalling pathways, whilst simultaneously enhancing susceptibility to the fungal pathogen *Alternaria brassicicola* via repression of chitin-mediated defence responses (Son *et al*., 2012). An example within *Brassicas* is *BrERF11* from *B. rapa* which provides enhanced tolerance to *Ralstonia solanacearum* infection when constitutively expressed in transgenic tobacco plants (Lai *et al*., 2013). This induces immune responses which include HR, oxidative bursts and upregulation of the pyhtohormones jasmonic acid (JA), salicylic acid (SA) and ethylene (ET) responses in a manner that suggests *BrERF11* is involved in PAMP or ETI.

Characterisation of the candidate gene Bo2g016500 is hampered by lack of annotation**,** though its homologue in *Arabidopsis* (AT5G18460) is described as a carboxyl-terminal peptidase (CTP). CTPs are ubiquitous in all three domains of life and have been associated with virulence in bacteria as a requirement for function of the type III secretion system. In plants, these protease enzymes are typically located within chloroplasts and are involved in C-terminal processing of the chloroplastic D1 protein of photosystem II. Proteolytic processing by *Arabidopsis CTPA3* for example facilitates light-driven assembly of the tetranuclear manganese cluster and has been described on UniProt (UniProt Consortium, 2021) as necessary for photosynthetic water oxidation.

Both Bo2g016490 and Bo2g016500 were confirmed by WGS to be monomorphic, though each contains EBH527 mutations within upstream UTRs, which could be affecting promoter activity. The role of promoter organisation in plants for determining transcriptional regulation of complex gene networks in response to pathogen invasions is becoming increasingly well documented, particularly with regards to determining downstream immune responses (Smirnova & Kochetov, 2015). Mutational disruption of host plant promoters can induce loss-of-function of *S*-genes, though documented examples seemingly all involve induced resistance to bacterial pathogen *Xanthamonas spp*., where transcription activator-like (TAL) effectors are prevented from recognizing their effector binding element (EBE) targets within promoter regions. Examples of this

includes the Lateral Organ Boundaries Domain gene *CsLOB1* in grapefruit variety Duncan, where promoter disruption using CRISPR-Cas9 induces resistance to *X. citri ssp. citri* (*Xcc*). Mutational knockouts of the *OsSWEET14 S*-gene promoter in rice also show subsequent resistance to bacterial blight (*X. oryzae pv. oryzae*) (Blanvillain-Baufumé *et al*., 2017). However, literature searches conducted here found no examples where *S*-gene promoter mutations determine resistance to other pathogen classes.

The remaining candidate Bo2g016510 is a G-type lectin S-receptor-like serine/threonine-protein kinase (STPK), which are typically localised to the cell wall or extra cellular matrix (ECM). This is the same class of *R*-gene represented by *Pto* which interacts with *Pseudomonas syringae pv. Tomato* effector *AvrPto* to confer resistance in tomatoes (Ronald *et al*., 1992). STPK domains play a well-established role in plant immunity, facilitating signal transduction pathways involving phosphorylation cascades upon *R*-gene-mediated detection of the pathogen (X. Wang *et al*., 2013). The Bo2g016510 homologue in *Arabidopsis* is At5g18470, a Curculin-like (mannose-binding) lectin family protein, involved in the specific recognition of mannose-containing glycans (Barre *et al*., 2019). In pepper, transient expression of the homologue *CaMBL1* has been found to induce accumulation of salicylic acid (SA), activating defence-related genes and triggering HR when tested with bacterial and fungal pathogen classes (Hwang & Hwang, 2011). Expression analysis in *A. thaliana* has also shown significant increases in At5g18470 regulation after independent inoculation with the oomycete *Phytophthora infestans* (after six hours) as well as the fungus *Erysiphe orontii* (after five days) relative to water controls (*Arabidopsis* eFP browser 2.0). At5g18470 was also listed as a significantly expressed gene in *Arabidopsis* (Col-0) when infected by the hemibiotrophic fungus *Colletotrichum higginsianum* (Hwang & Hwang, 2011).

The discovery from WGS collected here of three EBH527 SNP mutations within CDS of the single Bo2g016510 annotated intron highlighted it as a likely candidate, however each were found to be synonymous in their effect on amino acid composition in the translated protein. The central dogma of biology (Crick, 1970), traditionally considers synonymous codon changes to be 'silent' in their effects on protein function. However, this viewpoint has changed significantly over the last decade in the light of new evidence. Evolutionary studies for example have shown synonymous mutations to be under

selection pressure, a finding that runs counter to the established idea that they have a neutral effect on organisms fitness (Plotkin & Kudla, 2011).

A phenomenon called codon bias has been described whereby synonymous codons are utilised at different frequencies by translation machinery to regulate both the speed and extension length of the polypeptide chain (Sharp *et al*, 1993). Positive selection can thus act on synonymous variants that increase translation efficiency in order to tailor optimal protein yield. Codon bias is now recognised as critical for modulating gene expression and cellular function through impacts on RNA processing, protein translation and folding. Results from human studies that utilise large GWAS datasets have attributed at least 50 human diseases to associations made with synonymous mutations (Sauna & Kimchi-Sarfaty, 2011). In plants however, studies on the effects of non-neutral synonymous mutations with regards to disease resistance remains an understudied area of research.

One can conclude that if ACA2 resistance is determined by one of the three remaining candidate genes then it would represent a novel mechanism for determining susceptibility in plants, determined either by transcriptional regulation (in the case of Bo2g016490 or Bo2g016500) or via effects of synonymous mutations (in the case of Bo2g016510). To proceed with testing of candidates it would firstly be recommended to re-test the flanking markers used to define the ACA2 interval, to ensure the physical positions are correct. Secondly, it would be necessary to perform independent loss-of-function experiments on each of the remaining three candidates.

Even though knockouts of the GDSL-Lipase Bo2g016480 produced for this study provide strong evidence that the gene is not a susceptibility candidate, it is important to consider other genetic factors that may be influencing the result. For example, diploid *Brassica* species have undergone a whole genome triplication event since diverging from *Arabidopsis*, and multiple copies of genes can therefore be maintained. It is possible therefore that the CRISPR baits designed for this knockout are only targeting one functional copy of Bo2g016480 whilst other paralogues exist to facilitate pathogen infection. Another possibility is that there may be other genes in the ACA2 interval that are not detected here due to structural alterations in the parent lines relative to the reference assembly. For example, the ACA2 locus in *Arabidopsis* contains nine additional genes to that annotated within the *B. oleracea* TO1000DH3 reference. This demonstrates

an inherent limitation of referenced based mapping, which could be resolved by generating *de novo* assemblies of the EBH527 and A12DH parents.

The genetic background of the transformable host accession can also introduce uncertainty to the CRISPR results. Whilst DH1012 shares half its genetics with the A12DHd parent, including the Bo2g016480 allele, ideally the edit would be performed in A12DHd directly, where all other genetic factors are the same. Significant research efforts are now being put towards development of transformation methods for CRISPR/Cas transgene delivery in previously recalcitrant accessions (Sandhya *et al*., 2020). Here however, potential interactions due to the altered genetic background cannot be ruled out as influencing the result. For example, this recessive resistance may work epistatically with other genes, and where the second gene is not present (perhaps in the DH1012 background) then susceptibility is maintained. The identification and validation of the ACA2 gene has eluded efforts undertaken here, though now at least there is a better understanding of its complexities.

3.6.2 Objective 2 discussion. Elucidation of a dominant resistance gene tightly linked to ACA2 was an important discovery from the A12DH x EBH527 mapping population. Here, WGS-BSA was utilised to identify a highly significant QTL on chromosome 2 (Bol_EH177_C2) in progeny of line EH177. This was adjacent to the ACA2 locus and confirming tight-linkage of the two resistance loci. A second QTL on C6 (Bol_EH177_C6) also suggests activity of a second resistance locus in EH177. Dominant resistance in plants is primarily determined by NLR or RLK classes of *R*-gene and the EH177 QTLs support this as they both span TNL gene clusters (nine in Bol_EH177_C2 and two in Bol_EH177_C6). However, it remains unclear which of the TNL copies within these clusters is functioning in EH177 race 9 resistance.

Identification of EH177 resistance was fortuitous, due to the confirmed loss of ACA2 recessive resistance allele in this line which unmasked the activity of a secondary source of dominant resistance. Multiple layers of resistance are an important feature of white rust resistance in *A. thaliana.* The accession Columbia (Col-0), for example, contains several *WRR* genes that each provide a combined layer of resistance. The first of these to be mapped was the dominant, broad spectrum white rust resistance gene *WRR4* in *A. thaliana* accession Columbia. This TNL receptor induces a rapid defence

response that arrests pathogen development in the first epidermal cell, effectively masking phenotypes conferred by other resistance genes in Col-O. These were later confirmed as a pair of TNLs (*WRR5* and *WRR5b*) that induce a yellowing host response (Cooper, Cevik & Holub, unpublished), as well as a small LIM domain protein (*WRR7*) which exhibits flaccidity of the cotyledon and minor sporulation (Holub & Cevik, unpublished). A similar discovery was made when studying natural pyramiding of non-host resistance (NHR) genes in *A. thaliana*, where four genes (paralogues *WRR4a* and *WRR4b*, *WRR8*, *WRR9*, *WRR12*) were discovered that contribute to NHR of race 2 isolate Ac2v (Cevik *et al.*, 2019).

Like *WRR4*, it is possible that the broad spectrum ACA2 resistance acts early in arresting pathogen development, precluding any possible activation of EH177 defence in resistant parent EBH527. This is supported by the likelihood of the Bol_EH177_C2 resistance gene acting intracellularly as an *R*-gene receptor, whilst ACA2 candidates are more likely (based on the GDSL lipase Bo2g016480 and STPK Bo2g016510 candidates) to be located within the extracellular matrix and therefore activated first. It would be useful for future work to assess this by studying the relative advancement of the pathogen within EBH527 resistant (ACA2) vs EH177 resistant tissue, using techniques such as trypan blue staining.

The location of the highly significant Bol_EH177_C2 QTL, in close proximity (~ 66 kb) to the previously mapped recessive ACA2 locus was a key finding from this study, raising the question of whether there is an interactive relationship between the two loci. Whilst EH177 resistance has been shown to function independently of ACA2 resistance, it is unknown if EH177 is a requirement of ACA2 resistance. Confirming any interactive relationship between the two loci is important, not only for potentially characterising of a novel resistance complex in *Brassicas*, but also for practical efforts to identify the ACA2 causal gene. For example, loss-of-function experiments of ACA2 candidates in the transformable DH1012 line would be invalidated as this background lacks the Bol_EH177_C2 resistance allele, which would then maintain the GDSL-Lipase Bo2g016480 as the primary ACA2 resistance candidate. Alternatively, the co-localisation of the two traits may be coincidental, with the only currently available evidence to suggest otherwise being their close physical proximity.

Clustering of *R*-genes due to local duplication events is a typical feature of plant genome evolution, whereby variation can be generated and NLRs can be coregulated for effective triggering of immunity. Many *R*-genes/NLRs act in a singular gene-for-gene relationships with a pathogen substrate, whilst others require additional NLR proteins to function, forming connections that vary from pairs to complex networks (Adachi *et al*., 2019). Studies that demonstrate resistance mechanisms involving activity of at least two NLR predominantly show close genetic linkage between the loci (Ashikawa *et al*., 2008; Birker *et al*., 2009; Narusaka *et al*., 2009; Okuyama *et al*., 2011). Recessive resistance in barley to the wheat stem rust pathogen *Puccinia graminis* is a good example of this which requires the interaction of a complex of three tightly linked genes including two NLRs and a non-NLR (actin depolymerizing factor-like gene, *Hv*Adf3) (Wang *et al*., 2013). This common requirement of two or more *R*-proteins in a resistance complex, coupled with high modular variation in NLR genes, indicates that pathogen recognition is often dependent on different combinations of domains from distinct proteins.

The guard hypothesis lends support to this concept, where an *R*-protein (guard) can act by monitoring the target of the corresponding pathogen effector (guardee), inducing a rapid defence response upon modification of the target (Van Der Biezen & Jones, 1998). Interestingly, STPK proteins (such as Bo2g016510) are understood through well documented examples like *Pto* resistance in tomato to play a role as guardee proteins that mimic effector virulence targets (Ntoukakis *et al*., 2014). Resistance in *Arabidopsis* against *Pseudomonas syringae* strains carrying *AvrPphB* requires two genes; a NLR (*RPS5*) and an STPK (*PBS1*) (Swiderski & Innes, 2001). *PBS1* is constitutively bound by the guard *RPS5*, which detects AvrPphB-specific cleavage of the *PBS1* effector target to trigger an immune response (Van Der Hoorn & Kamoun, 2008). It is possible that the ACA2 resistance works in such a manner in combination with a local NLR guard protein (potentially contained within the Bol_EH177_C2 locus). However, without firstly identifying the ACA2 resistance gene this hypothesis cannot be tested and so remains as speculation.

With or without elucidation of the ACA2 resistance mechanisms, the discovery of the dominant resistance locus in EH177 only enhances the potential value of EBH527 resistance for utility in *Brassica* crops. Its tight-linkage to ACA2 provides natural pyramiding of the two loci for efficient MAS breeding into commercial cultivars, where

pathogen strains that overcome ACA2 resistance would also need to evade detection by the EH177-resistance protein to enhance durability of the trait. This could be achieved by selecting for the EBH527 haplotype across the combined EH177 and ACA2 region to ensure successful introgression of both loci for increased durability of the resistance.

Chapter 4

Pathogenomics of *Albugo candida* race 2 isolates from English mustard production in the UK

4.1 INTRODUCTION

Plant pathogens are continuously evolving in response to the selection pressure exerted by disease resistance in the host, which can drive 'boom-and-bust' cycles of resistance crop varieties that succumb to sequential proliferation of virulent pathotypes. Anticipatory breeding of disease resistance is a strategy which requires regular monitoring of pathogen populations to identify virulent (resistance-breaking) pathotypes before they become prevalent, which can then be used to screen germplasm resources for mapping and deploying the next resistance genes in a new crop variety (McIntosh & Brown, 1997).

Anticipatory resistance breeding (ARB) is predicated on the likelihood that several years may pass between detection of a new virulent pathotype and the occurrence of a loss causing epidemic, providing the necessary time for breeders to select and enhance recognition specificities within cultivars. The following are requirements for effective ARB: 1) knowledge of the pathogen epidemiology across the cropping system; 2) regular surveys across the cropping system aimed to detect emergent virulent strains that break currently deployed resistance; 3) knowledge of the main resistance genes currently deployed or in development; and 4) a well-coordinated system for screening germplasm using pathotypes that provide the greatest threat. Resistance genes identified in this manner target the pathotype that breaks currently established resistance and are therefore priority gene targets for introgression into advanced cultivars.

Advances in pathogen genomics (pathogenomics) can potentially support ARB by characterising variation in pathogen effectors that correspond with proteins encoded by matching *R*-genes. Once a putative repertoire of effectors has been identified, each can then be transiently expressed in a host to identify corresponding *R*-alleles that trigger an immune response (HR). Insights gained through the culmination of such effector targeting strategies can ultimately inform breeders when selecting the optimal combination of *R*-alleles in a breeding program. These can then be developed for introgression into a crop either through transgenic 'stacking' of alleles into a DNA construct for GM, or using conventional marker assisted selection to pyramid resistance alleles from several loci.

For example, the *Albugo candida* - *Brassica* pathosystem has been used extensively to investigate adaption of a pathogen across different closely related host

species. Host specialisation within *A. candida* has produced many distinct subspecies or physiological races (Jouet *et al*., 2019), which were previously classified according to their pathogenicity on a set of differential lines that represent different host species (Saharan *et al*., 2014, Srivastava *et al*., 2004, Hill *et al*., 1988, Pound & Williams, 1963). Race 2, for example, causes severe annual losses of oilseed mustard (*Brassica juncea*) in India, Canada, and Australia. This pathogen has emerged recently as a pathogen in UK English mustard production (Holub and Vicente, unpublished). Races 2 and 7 (derived from *B. rapa*) can be outcrossed under laboratory conditions (Adhikari *et al.,* 2003). Comparative genomics has also confirmed extensive introgression between races 2, 7 and 9 (McMullan *et al*., 2015), representing an approximate physical exchange of 25% of each genome. Consequently, such recombination amongst races could potentially spawn new variants with novel effector allele repertoires that facilitate expansion onto new hosts.

The original reference assembly of Canadian isolate of *A. candida* race 2 (Ac2vRR) revealed a compact genome of 45.3 Mb, which contained an estimated 15,824 genes (Links *et al*., 2011). The genome displays a signature of obligate biotrophy having lost certain key enzymes typically present in phytopathogen genomes, such as Necrosis and Ethylene inducing Peptides (NEPs). A unique class of secreted effector candidates has also been determined, possessing a CxxCxxxxxG amino-acid motif (abbreviated as CCG) that are able to translocate into plant cells to supress host immunity (Kemen *et al*., 2011). Each race of *A. candida* has been documented to contain ~100 CCG secreted proteins that show considerable presence absence polymorphism (Redcar *et al*., unpublished 2021, Jouet *et al.*, 2019). The Ac2vRR genome assembly has recently been improved with the use of Pacific Biosciences (PacBio) long read data and Illumina short read augmentation (named Ac2VPB), increasing the assembly size by 10% (Furzer *et al*., submitted 2021). Additional RNAseq data was used to improve CCG annotation which increased the number of predicted CCGs by 250%, suggesting that other Illumina based *Albugo spp.* assemblies will be underestimating presence of CCGs.

Comparative genomics enables the identification of shared or isolate-specific pathogenicity factors. This process has been enhanced by long-read (third generation) sequencing platforms which make it possible to rapidly sequence and assemble high-quality pathogen genomes. It is essential that a contiguous genome assembly is acquired

to identify and characterise pathogenicity factors such as effectors, which frequently reside in highly repetitive and rapidly evolving genomic regions (Thomma *et al*., 2015).

Furthermore, a contiguous assembly can improve accuracy in the prediction of genes and co-regulated gene clusters, which have found to play a role in pathogenicity (Bolton et al., 2014). Oxford Nanopore Technologies (ONT) provides a modern long-read sequencing platform that enables the generation of high-quality genome assemblies at affordable costs. The portable MinION sequencer has already played a key role in monitoring numerous infectious disease outbreaks (Jain *et al*., 2016; Mongan *et al*., 2020), allowing field isolates to be multiplexed within a single chip for real-time sequencing and comparative genomics.

The *A. candida – B. juncea* pathosystem is conducive to an effector targeting strategy for anticipatory resistance breeding, as milestones for achieving this have already been met. Firstly, race 2 isolates have already been identified in India that break the only identified source of WRR in *B. juncea* (*BjuWRR1*) (Dev *et al*., 2020), as well as Canadian isolate Ac2v and the UK isolate AcBjDC (established here in Section 2.3.1). The *BjuWRR1*-virulent UK isolate AcBjDC has been used to identify a new WRR locus in *B. rapa* (Chapter 2) designated as EH_25_DC.A06, providing an additional target for combining with *BjuWRR1* in a breeding program. What remains as a prerequisite for ARB is to document the comparative effector repertoires of resistance breaking isolates. This will be pursued here by utilising ONT long-read sequencing to compare two UK isolates (AcBj12 and AcBjDC) with Canadian isolate Ac2V.

4.2 METHODS

4.2.1 Plant and pathogen maintenance. The *B. juncea* cultivar Sutton was used to propagate two UK isolates of *A. candida* race 2 (AcBj12 and AcBjDC) in the manner described previously in Section 2.2.1. Three trays containing approximately 180 seedlings were used per isolate to produce sufficient quantities of spores required for DNA extraction. Donskaja was included as a control in each tray as a means of distinguishing the two isolates by their varying degree of sporulation (see Figure 2.2).

4.2.2 DNA extraction and quality control. Zoosporangia were harvested from heavily infected cotyledons using a handheld cyclone spore collector. Spores were ground to a fine powder using a pre-chilled pestle and mortar with liquid nitrogen. DNA extraction was carried out using the CTAB method as described in Section 2.2.4, though taking extra precautionary steps to ensure the DNA was not inadvertently fragmented. Care was therefore taken once the tissue was lysed, which included widening the aperture of pipette tips with a sterilised scalpel as well as extremely slow pipetting of the DNA in solution. Extracted DNA samples were visualised on a 1% agarose gel to check for the occurrence of a discrete and narrow high-molecular weight band without presence small DNA fragments. DNA yields were confirmed using a Qubit dsDNA BR Assay Kit with a Qubit 2.0 Fluorometer (Invitrogen, Waltham, USA) to confirm sufficient DNA was obtained per sample (≥ 1 ug). DNA quality was assessed using a NanoDrop Spectrophotometer (NanoDrop, Wilmington, USA) to ensure a 260/280 ratio between 1.8 and 2.0 and a 260/230 ratio between 2.0 and 2.2 was achieved.

4.2.3 MinION sequencing library preparation. The online Oxford Nanopore protocol – native barcoding genomic DNA (with EXP-NBD196 and SQK-LSK108) was followed for library preparation of high-molecular weight DNA. Samples were prepared for a multiplexed single flowcell sequencing run, including DNA from AcBj12, AcBjDC and an *A. candida* race 4 isolate AcEM2 (whose data were not assessed further in this study). Library preparation required an initial DNA end repair step using a mastermix of NEBNext FFPE DNA Repair Mix (M6630) NEBNext Ultra II End repair / dA-tailing Module (E7546) reagents, combined with 400ng of DNA per sample. This was incubated at 20°C for 5 mins

and 65°C for 5 mins. End-prepped DNA was then used for ligation of native barcodes, with long-fragment buffer applied to enrich for fragments > 3 kb. This buffer was used along with Agencourt AMPure XP beads to pellet, elute and purify the DNA from residual reagents. MinION sequencing was performed using a FLO-MIN107 R9 flowcell, MinKNOW 1.15.1 and a standard 48-hour run script with active channel selection enabled. Reads were base called in real time on-instrument using the Guppy v2.2.2 GPU basecaller. Sequencing summary statistics are presented in Table 4.1.

4.2.4 Read processing and *de novo* assembly. Raw fast5 and fastq ONT read data was given to Dr Volkan Cevik at the University of Bath for bioinformatic processing and construction of genome assemblies, which were then returned for all remaining analyses. Initially, raw fast5 sequence reads for the two *A. candida* isolates were demultiplexed using Deepbinner (v0.2.0, (Wick *et al*., 2018)) before removal of adapter and index sequences from fastq files using Porechop (v0.2.1, (Wick *et al*., 2017)). These were then converted to fasta format before undergoing read correction using LoRDEC (v0.6, (Salmela & Rivals, 2014)) and previously generated Illumina short-read MiSeq data of the sequenced isolates. The corrected reads were then used to generate polished *de novo* assemblies using Flye (v2.8.2, (Chen *et al*., 2020)). To account for errors in the assembly, three rounds of PILON polishing were conducted (Walker *et al*., 2014), using BWA-MEM for mapping of Illumina reads (Li & Durbin, 2009). Complete assemblies were then recieved by Dr Cevik to the author, where final assembly statistics were generated for both ONT assemblies as well the PacBio Ac2vPB assembly using QUAST (v5.0.2, (Gurevich *et al.*, 2013)) (Table 4.2). This includes an estimate of genome completeness, conducted using BUSCO (v3, (Simão *et al*., 2015)) to search for near-universal single-copy fungi orthologues in the three assemblies.

4.2.5 Prediction of open reading frames (ORFs) and CCGs. ORFs were determined from the AcBj12 and AcBjDC ONT genome assemblies, as well as the Ac2v PacBio assembly provided by Dr Cevik using the online EMBOSS tool getorf (http://www.bioinformatics.nl/cgi-bin/emboss/getorf), which translated regions between start and stop codons from forward and reverse sequence. A list of annotated CCGs identified from previous genome projects (Jouet et al., 2019; Links et al., 2011; McMullan

et al., 2015) was then used as a template for generating a CCG motif with MEME (v5.1.1, (Bailey *et al*., 2009)). This motif was then used to screen translated ORFs from the three *A. candida* assemblies using MAST (Bailey *et al*., 2009; Bailey & Gribskov, 1998) to identify those containing probable CCG motifs in our datasets (E-value <0.1, correct positioning for the CCG motif). The output of probable CCG proteins was then filtered based on the presence of a signal peptide as predicted by SignalP-5.0, excluding any with SP signal of <0.01% (Almagro Armenteros *et al*., 2019). A comparison of predicted CCG protein sequences between isolates was then conducted with a combination of BLAST and USEARCH v11 (Edgar, 2010) (Table 4.3).

4.3 RESULTS

MinION sequencing of *Albugo candida* isolates AcBj12, AcBjDC and AcEM2. High-molecular weight DNA was extracted from *A. candida* race 2 isolates AcBj12 and AcBjDC and was submitted for sequencing on the Oxford Nanopore MinION platform. A third isolate AcEM2 was also included in the library multiplex, though is not part of further analysis here. In total 3.53 Gb of read data was generated for the three ONT sequenced isolates, which equates to an average coverage of the *A. candida* genome of 22.5x per isolate and an average read length of 4012 bp (Table 4.2). Whilst this produced sufficient data to proceed with assemblies, both the total yield and fragment lengths produced were significantly under the design capacity of the flowcell used. Productivity of the flowcell was impeded by a low availability of functional pores, whilst fragment length was limited due to shearing that occurred during library preparation.

Table 4.4. Summary statistics for MinION sequencing data after completion of a 48-hour run from three *Albugo candida* isolates (AcBj12, AcBjDC and AcEM2). High-molecular weight DNA from were multiplexed and run on a single FLO-MIN107 R9 flowcell.

	AcBj12	AcBjDC	AcEM2
Gigabase pairs (Gb)	1.08E+09	1.52E+09	4.54E+08
Number of reads	267150	424347	102697
Mean read length	4036	3583	4419
Mean coverage (x)	23.8	33.6	10.0

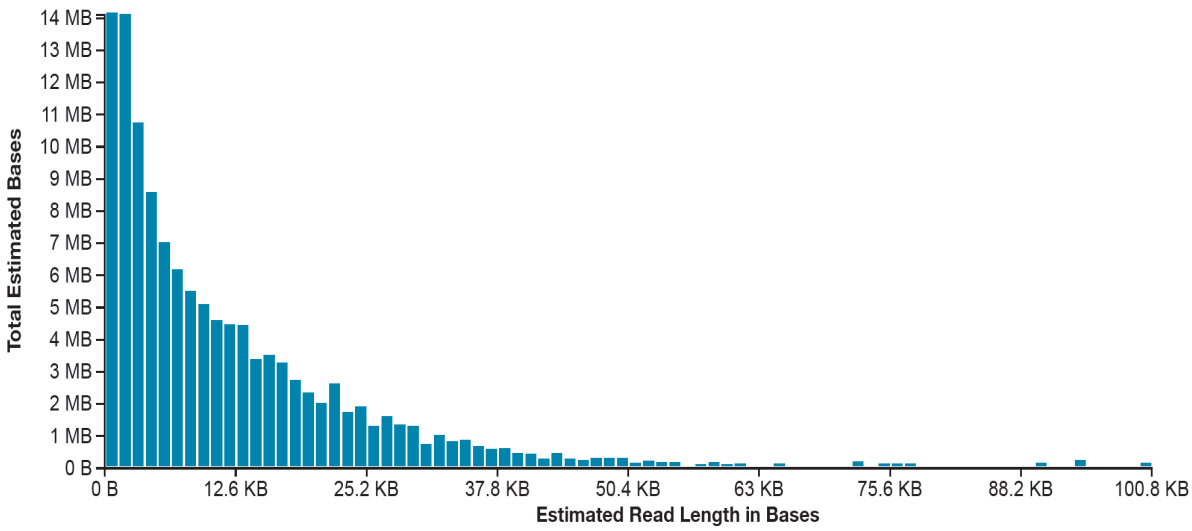


Figure 4.1. Read length of MinION sequencing run that contained a multiplexed DNA library of the three *Albugo candida* isolates AcBj12, AcBjDC and AcEM2, extracted at 12 hours into the 48-hour run. Whilst some high-molecular weight DNA fragments were sequenced these results indicate that substantial shearing occurred during library preparation.

4.3.2 *De novo* assembly of AcBj12 and AcBjDC genomes from MinION read data. Prior to assembly, raw Oxford Nanopore (ONT) reads were corrected with LoRDEC (v0.6, (Salmela & Rivals, 2014), using previously obtained Illumina MiSeq reads from isolates AcBj12 and AcBjDC. Corrected reads then underwent *de novo* genome assembly using Flye (v2.8.2, (Chen *et al*., 2020), before three rounds of Pilon polishing with a BWA-MEM (algorithm currently unpublished) Illumina read alignment. The two ONT assemblies have been named by corresponding isolate as AcBj12_ONT and AcBDC_ONT.

When considering that both the yield and average fragment lengths sequenced here were considerably below the capacity of the ONT flowcell, the two assemblies produced are none-the-less closely comparable to the high-quality PacBio Ac2vPB assembly (Furzer *et al*., unpublished 2021). Assembly statistics in Table 4.2 show that Ac2vPB is slightly larger than the two ONT assemblies in terms of total alignment length by an average of 4.36 %. However, the AcBj12_ONT assembly in particular demonstrates an improvement on contiguity with an N50 of approximately 630 kb (29.3 % greater than that of Ac2vPB and 57 % greater than the original Illumina assembly Ac2vRR). Relative to the Ac2vRR Illumina assembly that contained 1.7 million Ns across unresolved repetitive regions, the ONT assemblies contain relatively few at an average total count of 2693, whilst the Ac2vPB assembly contains only 100. Structural comparisons of the genomes,

using the online tool D-Genes (Cabanettes & Klopp, 2018) displays a high degree of similarity between the three assemblies, particularly between AcBj12_ONT and AcBj12_BJ (Figure 4.5). Comparisons made between either ONT assembly and Ac2vPB display a greater number of INDELs, as well as a shared structural inversion of a single 475 kb scaffold (Figures 4.3 and 4.4).

Table 4.5. Genome assembly statistics for an improved reference genome of *Albugo candida* race 2 using long-read PacBio sequencing (referred to as Ac2vPB), and comparative *de novo* assemblies of DNA from two UK isolates using Oxford Nanopore Technology (AcBj12_ONT and AcBjDC_ONT). Presence of complete single-copy BUSCOs were established in each assembly using the fungi lineage dataset fungi_odb10. These provide a proxy for relative completeness of an assembly.

Genome statistics	Ac2vPB	AcBj12	AcBjDC
Genome fraction (%)	93.84	88.85	89.40
Total aligned length (bp)	37372940	35540618	35945743
No. contigs	198	230	376
N50	466138	657574	491022
L50	29	21	25
Mean contig length (kb)	196	160	98
Largest contig (bp)	1218497	1420159	1130785
N's per 100 kbp	0.26	8.67	5.88
Complete fungal BUSCOs number (%)	361 (47.5)	364 (48)	362 (48)

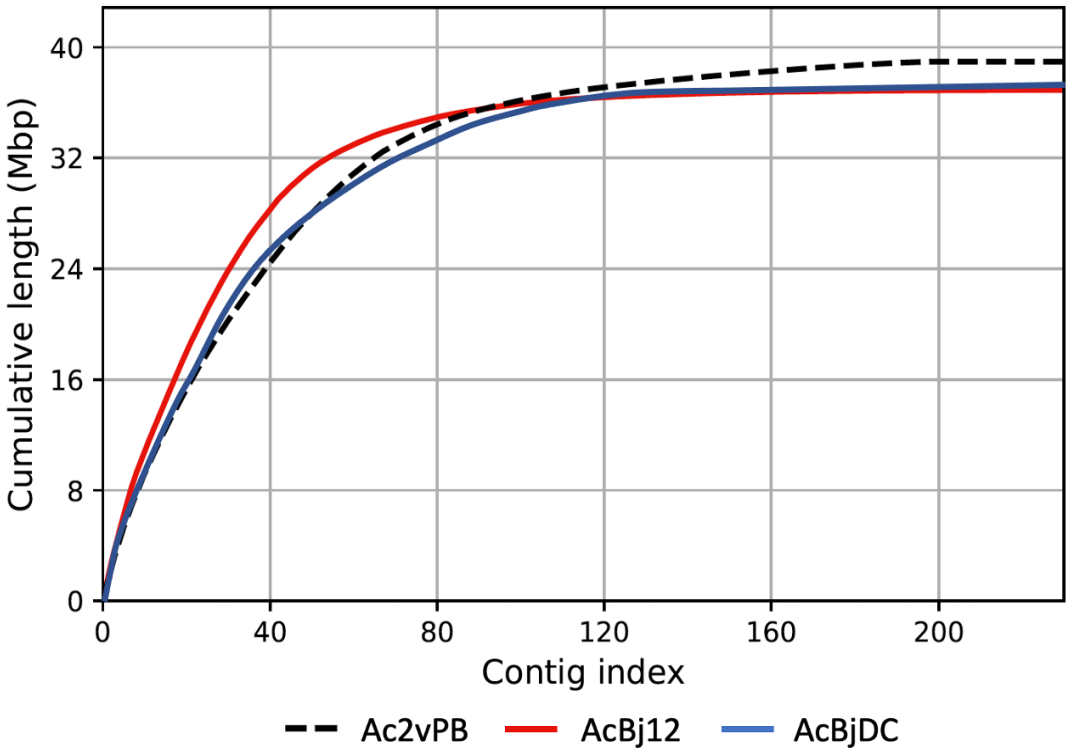


Figure 4.2. Plot comparing the cumulative increase in genome sizes of size-sorted contigs for three genome assemblies of *Albugo candida* race 2, including Ac2vPB (PacBio), AcBj12_ONT and AcBjDC_ONT (ONT). Saturation in the increase of the assembly lengths can be seen here within the first 240 contigs presented. The AcBj12_ONT assembly can be seen to displays the most efficient genome construction, assembling the largest proportion of the genome using the least contigs. Whilst the Ac2vPB has a better completeness (assembled length) of the assembly. Give a key to the different coloured lines

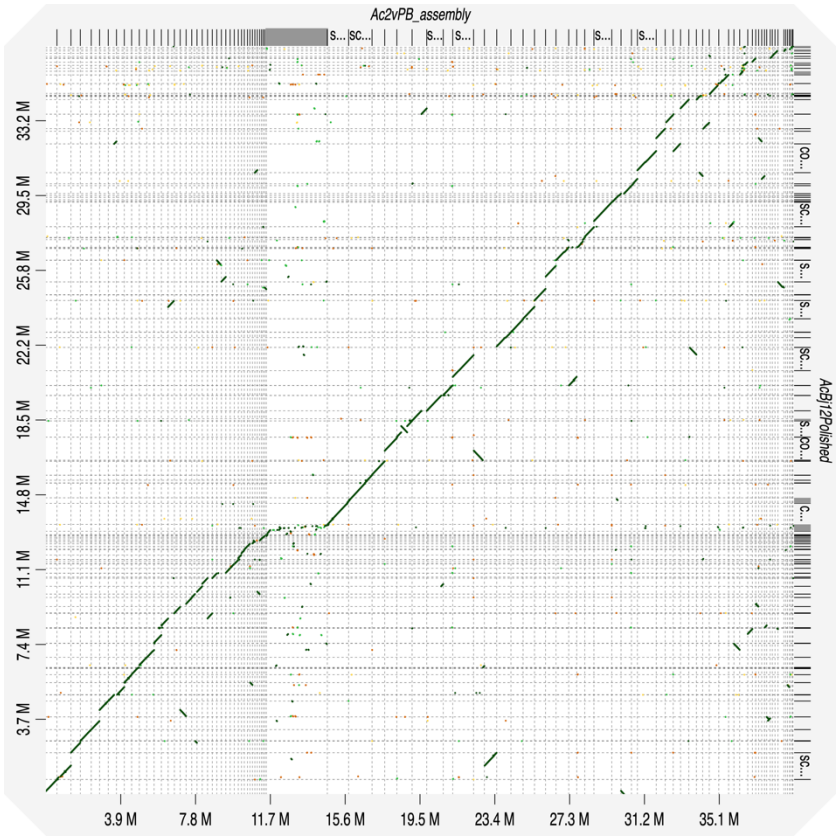


Figure 4.3. Dot plot comparison of the PacBio Ac2v assembly (Ac2vPB) versus the ONT AcBj12 assembly (AcBj12_ONT). The graphic was produced using the online tool D-Genies, which uses Minimap (v2) to align the two genomes and provide a synthetic similarity overview. Features highlighted include repetitions (lines parallel to the diagonal line), breaks (deletions, insertions and out of frame mutations) and inversions (lines perpendicular to the diagonal).

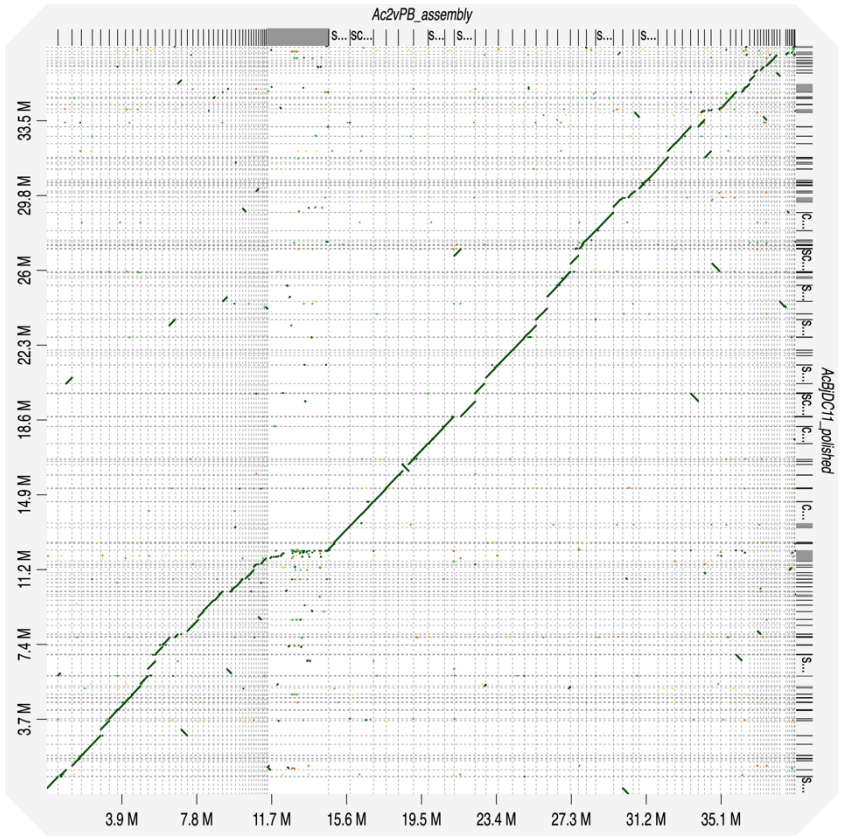


Figure 4.4. Dot plot comparison of the PacBio Ac2v assembly (Ac2vPB) versus the ONT AcBjDC assembly (AcBjDC_ONT).

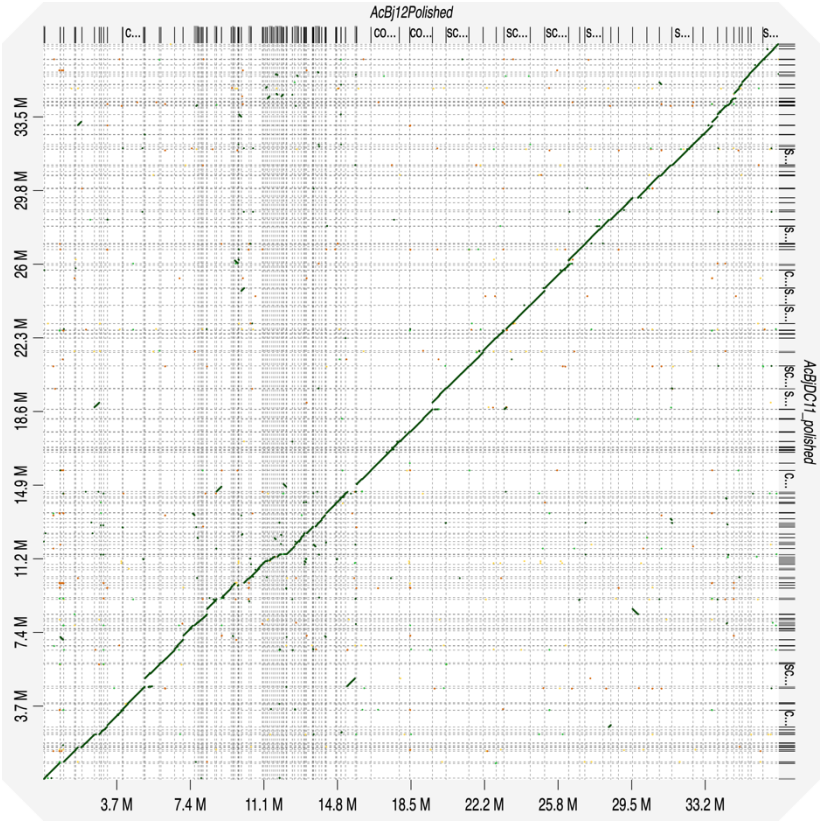


Figure 4.5. Dot plot comparison of the two ONT assemblies AcBj12_ONT and AcBjDC_ONT.

4.3.3 Search for candidate effector encoding genes. As CCG effectors are known to be encoded by single exon genes (no intronic regions), the process of identifying them from the three assemblies (Ac2vPB, AcBj12_ONT and AcBjDC_ONT) began by extracting all open reading frames (ORFs) with the online EMBOSS tool getorf. A list of previously identified CCGs from Ac2v assemblies Ac2vRR and Ac2vPB (supplied by Dr Cevik) was then used to construct the MEME motif in Figure 4.6 (Bailey *et al.*, 2009) which was used as the input for a MAST search of the ORF files generated from the three assemblies to identify likely CCGs (Steuernagel *et al.*, 2015). This list was then screened in SignalP-5.0 (Almagro Armenteros *et al.*, 2019), where any candidate with a SP signal < 0.01 was rejected to produce a final shortlist of ORFs containing both a detected CCG motif and secretion signal. From this search, a total of 133 CCGs were identified in Ac2vPB, 122 in AcBj12_ONT and 126 in AcBjDC_ONT.

A protein BLAST was conducted between these candidates and the list of previously identified CCGs, with results summarised in Table 4.3. Of the 128 previously annotated CCGs, exactly 100 of them were recognised in any of the three assemblies (sequence length and identity similar by > 90 %), with 96 found in Ac2vPB, 85 found in AcBj12_ONT and 92 found in AcBjDC_ONT. Of these, 33 annotated CCGs were fully conserved between the three assemblies whilst the remaining 67 showed some degree of variation. Examples were also found where the BLAST criteria matched more than one ORF, suggesting presence of a second paralogue.

When comparing candidate CCG alleles using USEARCH (v11, (Edgar, 2010)), a total of 44 proteins remained (collectively between isolates) that did not match (> 90 % similarity) to any annotated CCG. Specifically, this was 34 examples in Ac2vPB, 38 in AcBj12_ONT and 40 in AcBjDC_ONT. These additional CCG/signal peptide candidates have been compared between the assemblies and are also documented in Table 4.3 (examples that lack a "CCG" annotation/query ID).

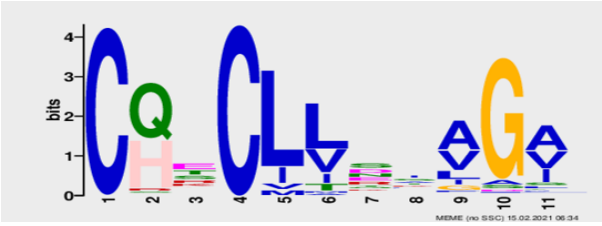


Figure 4.6. CxxCxxxxxG amino-acid motif (CCG) predicted in *Albugo candida* Ac2v effector proteins using MEME. Utilised as a search pattern for identifying similar proteins in the genome assemblies Ac2vPB, AcBj12_ONT and AcBjDC_ONT.

This table needs a number and title

Group	No.	Query sequence ID	Length (AA)	Ac2v	AcBj12	AcBjDC
Monomorphic	1	CCG51_Ac2v	115			
	3	CCG22_Ac2v	127			
	4	CCG62_Ac2v	127			
	6	CCG26_Ac2v	133			
	7	CCG13_Ac2v	145			
	8	CCG72_Ac2v	155			
	9	CCG49_Ac2v	168			
	11	CCG103_Bp135 _Ac2v	191			
	13	CCGlike41	213			
	15	CCG111_Vnew_PacBio_Ac2v	229			
	18	CCG1_Ac2v	258			
	19	CCG105_Vnew_PacBio_Ac2v	259			
	22	CCG65_Ac2v	277			
	24	CCGlike10	315			
	33	CCG3_Ac2v	559			
	36	CCG14_Ac2v	572			
	38	CCG55_HB_Ac2vRR_Ac2v	584			
	39	CCG55_HB_Ac2vRR_Ac2v	585			
	40	CCG44_Ac2v	590			
	43	CCGlike17_Ac2v	594			
	44	CCG70_Ac2v	598			
	47	CCG61_Ac2v	601			
	48	CCG106_Vnew_PacBio_Ac2v	602			
	54	CCG108_Vnew_PacBio_Ac2v	619			
	57	CCG17_Ac2v	625			
	61	CCG110_Vnew_PacBio_Ac2v	630			
	64	CCG66_Ac2v	640			
	75	CCG69_Ac2v	688			
	76	CCG36_Ac2v	691			
	82	CCG15_New_Ac2vRR_HB	708			
Monomorphic (continued)	86	CCG23_Ac2v	716			
	87	CCG8_Ac2v	716			
	89	CCG75_Ac2v	719			
	34	CCGlike16	565	A1	A1	A1
	69	CCG7_Ac2v	682	A1	A1	A1
	78	CCG58_Ac2v	700	A1	A1	A1
	99	CCG60_Pac_New_Ac2v	775	A1	A1	A1
	101	BjDC__scaffold_128_382	92			
	104	BjDC__contig_344_63	102			

	105	BjDC__contig_77_34	103			
	106	BjDC__scaffold_171_253	103			
	107	BjDC__contig_77_251	105			
	108	BjDC__scaffold_171_277	106			
	109	BjDC__contig_218_201	111			
	111	BjDC__scaffold_171_618	120			
	113	BjDC__scaffold_171_123	126			
	114	BjDC__contig_123_238	128			
	115	BjDC__scaffold_212_260	128			
	118	BjDC__contig_39_1	141			
	119	BjDC__contig_123_321	144			
	120	BjDC__scaffold_195_143	146			
	122	BjDC__contig_223_140	160			
	123	BjDC__contig_129_30	168			
	124	BjDC__contig_21_39	180			
	125	BjDC__scaffold_147_779	181			
	130	BjDC__contig_172_55	257			
	132	BjDC__scaffold_147_136	311			
	137	BjDC__contig_77_287	486			
	143	BjDC__contig_7_266	731			
	141	BjDC__contig_7_106	676			
	2	CCG10_New_PacBio_Ac2v	125	A1/A2	A2	A2
	30	CCG30_Ac2v	550	A1/A2		
	112	CCG10_New_PacBio_Ac2v	125	A1/A2		
Ac2V	14	CCG9+C65:C861Like_new_PacBio_Ac2v	217			
	17	CCG94_Pac_New_Ac2v	247			
	46	CCGlike35	600			
	88	CCG24_Pac_New_Ac2v	717			
	25	CCG16_Ac2v	403		A2	A2
	28	CCG32_Ac2v	534		A2	A2
	35	CCGlike5	565		A2	A2
	55	CCGlike29_Pac_New_Ac2v	621		A2	A2
	91	CCG45_Ac2v	722		A2	A2
Ac2V	92	CCG4_Ac2v	724		A2	A2
(continued)	98	CCG5_Pac_New_Ac2v	770		A2	A2
	68	CCG100_New_PacBio_Ac2v	677			
	45	CCG2_Ac2v	600	A2		
	83	CCG43_New_PacBio_Ac2v	708			
	29	CCG28_Ac2v	544	A2	A1	A1
	102	BjDC__scaffold_171_27	96			
	128	BjDC__scaffold_128_492	215			
	116	Ac2v__scaffold149_29	129			

	138	Ac2v__scaffold123_147	539			
	126	Ac2v__scaffold_3_369	197		A2	A2
AcBj12	20	CCG83_Ac2v	261			
	23	CCG86_Ac2v	302			
	26	CCG76_Ac2v	443			
	41	CCG71_Ac2v	590			
	59	CCG48_Ac2v	628			
	16	CCG33_Ac2v	242		A2	
	32	CCG41_Ac2v	554		A2	
	58	CCG19_Ac2v	628		A2	
	62	CCGLike30_New_PacBio_Ac2v	630		A2	
	65	CCGlike23_Ac2v	644		A2	
	100	CCG79_Ac2v	797		A2	
	90	CCG85_B_Pac_New_Ac2v	720		A2	
	70	CCG88_Real_Ac2v	682	A1	A2	A1
	94	CCGlike31_New_PacBio_Ac2v	744	A1	A2	A1
	121	Bj12__contig_431_73	155			
	129	Bj12__scaffold_92_631	217			
AcBjDC	10	CCGlike12	190			
	96	CCG37_Ac2v	752			
	66	CCGlike18	645			
	71	CCG77_Ac2v	684			
	5	CCG12_Ac2v	132			A2
	12	CCGlike40	197			A2
	31	CCG31_Ac2v	551			A2
	49	CCG46_Ac2v	604			A2
	50	CCGlike45_Ac2v	605			A2
	67	CCG104_Vnew_PacBio_Ac2v	659			A2
	81	CCG57_Ac2v	706			A2
	95	CCG37_Ac2v	751	A2	A2	
	63	CCG109_Vnew_PacBio_Ac2v	635	A1	A1	A2
	103	Bj12__scaffold_125_703	97			A2
	117	BjDC__contig_386_56	140			
	127	BjDC__contig_317_10	200			
AcBjDC	134	BjDC__scaffold_264_162	410			
(continued)	136	BjDC__contig_498_13	424			
	131	BjDC__contig_123_138	284	A2	A2	
	133	BjDC__contig_237_177	314	A2	A2	
Polymorphic	53	CCG35_Ac2v	616		A2	A3
	73	CCG101_Bp125hom_Ac2v	685		A2	A3
	74	CCG102_Vnew_PacBio_Ac2v	685		A2	A3
	77	CCG74_Ac2v	695		A2	A3
	79	CCG34_Ac2v	702		A2	A3

97	CCG81_93_Pac_New_Ac2v	754		A2	A3
52	CCG67_Ac2v	612	A2		A3
60	CCG80_Ac2v	628	A2	A3	
27	CCG54_Ac2v	487	A2		
60	CCG80_Ac2v	628	A2	A3	
37	CCGlike37_Ac2v	577	A1	A2	A3
72	CCG101_Bp125_AcBoT	685	A1	A2	A3
80	CCG11 _Allele2_Ac2v	704	A1	A2	A3
84	CCG38_HB2vRR_Ac2v	710		A2	A3
42	CCG56_Ac2v	591		A1	A2
85	CCG85_A_Pac_New_Ac2v	715	A1/A2	A1	A2
56	CCGlike34_Ac2v	623	A1		A2
93	CCG53_Ac2v	730	A1	A2	
21	CCG112_Vnew_PacBio_Ac2v	277			A2
51	CCG82_Ac2v	610			A2
27	CCG54_Ac2v	487	A2		
139	Bj12__scaffold_92_624	607	A2		A2/A3
142	BjDC__contig_77_368	719	A2	A3	
140	BjDC__contig_117_125	628	A2	A3	
110	BjDC__contig_187_378	113		A2	
135	BjDC__contig_77_19	412		A2	
144	BjDC__contig_317_23	745	A2	A3/A4/A4	A1/A2

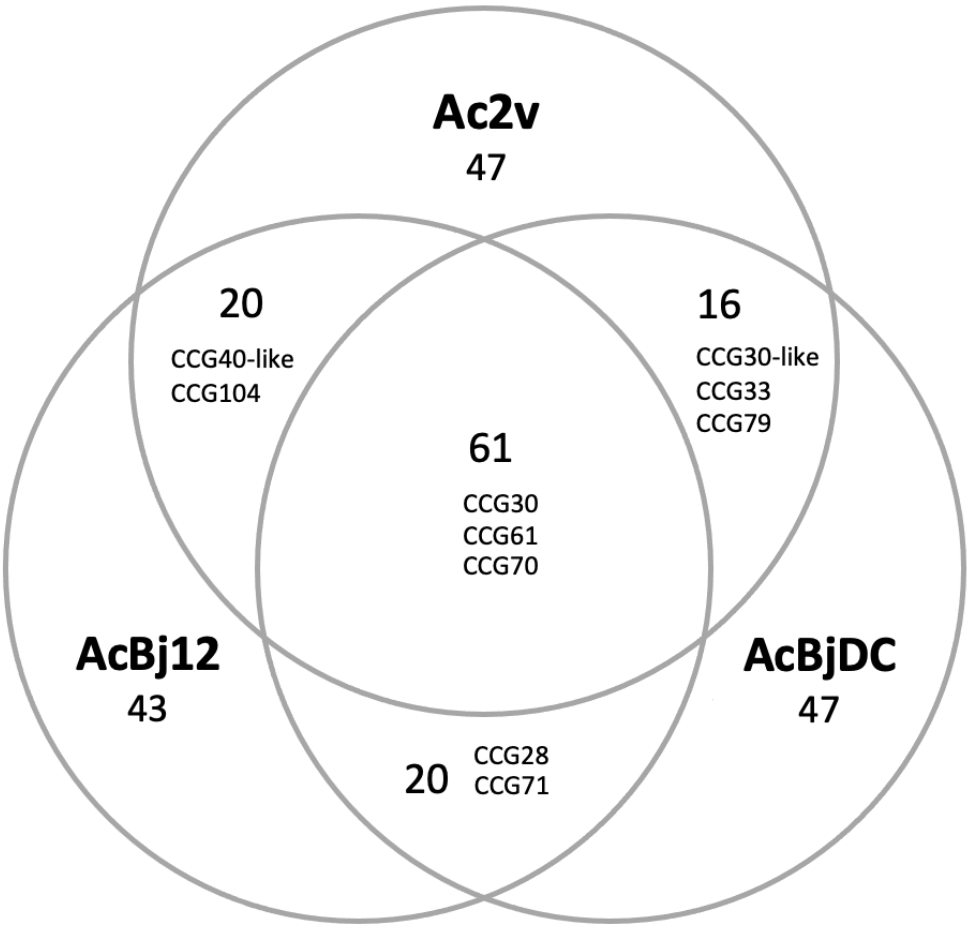


Figure 4.7. Ven diagram showing the relative overlap of CCG allelelic conservation between race 2 isolates Ac2v, AcBj12 and AcBjDC. Numbers reported where there is no overlap represent a count of unique alleles for the relative isolate. The annotated CCGs included here indicate examples of *Avr* elicitors that are recognised between the broad-spectrum white rust resistance alleles *WRR4A* or *WRR4B* alleles (Col-0), as decribed by Redcar *et al*. (unpublished 2021). Those included in the central overlap are important therefore as conserved effector targets, which would each need to be mutated for an isolate to avoid detection by *WRR4A* and *WRR4B*. This highlights the potential importance of including these *R*-alleles alongside *BjuWRR1* to ensure durable resistance in Indian *B. juncea* crops.

4.4 DISCUSSION

Oxford Nanopore Technology (ONT) provides an exciting new alternative for rapid DNA sequencing and assembly of high-quality pathogen genomes. This study reports the first known assemblies of an *A. candida* genome using this platform, with assembly statistics that closely match that of the best available reference genome of Canadian isolate Ac2v, which was recently updated using PacBio sequencing (Ac2vPB) (Furzer *et al*., unpublished 2021). Importantly, the AcBj12_ONT assembly improved genome contiguity by 23.4% from that of Ac2vPB. This was achieved from a sequencing run that was suboptimal in a single flowcell, so it is likely that further closure of the *A. candida* genome could be achieved in a second run.

The read length of nanopore sequences are theoretically only limited by the length of the inputted DNA molecules themselves. As such, this technology has the potential to generate ultra-long reads that span megabases in size, enabling the assembly and resolution of complex genomic regions. A recent human genome assembly, for example, maximised ultra-long reads (up to 882 kb) to yield a highly contiguous assembly of NG50 ~3 Mb from a MinION sequencer (Jain *et al*., 2018). For plant pathogens, nanopore sequencing of the ~54 MB *Rhizoctonia solani* fungal genome was used to generate an average read length of 10.7 kb, producing an assembly with an N50 contig length of 199 kb (Datema *et al*., 2016). However, this study took steps to actively fragment genomic DNA, to maximise coverage rather than contiguity.

Whilst the average read lengths of the current ONT assemblies are markedly lower than preferred, the AcBj12_ONT assembly was still able to provide a significant improvement on genome contiguity relative to Ac2vPB (N50 of 657574 vs 466138). Some reads were sequenced in the multiplex that were very long (> 80 kb), and the success of AcBj12_ONT may have been due to obtaining a greater ratio of these reads from the library. Capturing reads of this length likely aided assembly of a few very large contigs for this sample, as seen by the 1.42 Mb contig produced which represents the current longest *A. candida* contig assembled (3.14 % of the genome). However, an abundance of small contigs (< 5000 bp) in the ONT assemblies meant that it was not possible to reduce contig number relative to the Ac2vPB assembly. This suggests that improving both the

abundance and ratio of long reads (relative to short) in any nanopore assembly is key to improving contiguity.

Enhancements to nanopore flowcell chemistry and hardware continue to improve base calling accuracy of single reads, from what was originally 60% (Laver *et al*., 2015) to a recent account of 99.1% on a PromethION (Nanopore Community Meeting, 2020). At the time of sequencing this experiment however, and with the flowcell kits used, the accuracy of raw reads was found to be an issue. This was partly due to the poor performance of the flowcell, where low availability of active pores limited the output of sequence data. Consequently, the coverage of nanopore reads was insufficient for correcting sequence errors in the data analysis. Illumina short read data was therefore required to provide accurate base calling and construction of high-accuracy ONT assemblies.

Attainment of these high-quality genomes enabled identification of novel *A. candida* CCG class effectors in the two UK isolates AcBj12 and AcBjDC. Previous development of the PacBio Ac2v genome identified 110 CCGs, which provided over a two-fold improvement on numbers obtained from the Ac2v Illumina assembly Ac2vRR (Furzer *et al*., unpublished 2021). An updated list of 128 Ac2v CCGs was provided for this study by Dr Cevik, which is closely comparable to numbers gained from our assessment of Ac2vPB (133 CCGs) as well as AcBj12_ONT (122) and AcBjDC_ONT (126). Of these, 80% from Ac2vPB, 66% from AcBj12_ONT and 72% from AcBjDC_ONT sufficiently matched (> 90% length or sequence similarity) previously annotated Av2v CCGs (Table 4.3). Whilst an average of 28% do not closely resemble a previously identified Av2v CCG and could potentially represent newly identified effector proteins. However, this level of variation includes comparisons between Ac2v CCG alleles identified here and those identified from the previous study (using the same genome assembly), indicating that disparity between the two methods applied may be influencing selection of CCG alleles in each analysis.

Whether additional unique CCGs have been identified in this analysis or not, the majority documented here are established as this novel class of secreted *A. candida* effector and thereby provide a useful insight into effector variation between different strains of *A. candida* race 2. The extent of CCG diversity between AcBj12 and AcBjDC was unexpectedly high considering the two isolates were collected from the same UK field of *B. juncea* and share such similar genome architecture (Figure 4.5). As much as 42% of

their shared CCG effectors are variable between the two isolates, demonstrating the high mutation rate of these effector-encoding genes in the genome. This is even comparable to levels of CCG variation found between Canadian isolate Ac2v and both UK isolates, with 47% of their CCG effectors here being fully conserved.

Pathogenomic datasets such as the one provided here in Table 4.3 are useful tools for shortlisting a set of candidate *Avr* elicitors, where you have allelic effector sequence for both virulent and avirulent isolates. Candidates would need to fit an expected pattern of a conserved allele across avirulent isolates, that is mutated in virulent isolates (thereby avoiding detection by the corresponding *R*-protein). To progress with shortlisting Donskaja *Avr* elicitor candidates it would be necessary to extend the current dataset, predominantly with additional avirulent race 2 isolate data (suggested provision of at least two more isolates). This would strengthen the phenotypic association of any individual candidate CCG and provide a shortlist of effectors small enough to proceed with transient testing of alleles.

The recent study by Redcar *et al*. (unpublished 2021), used transient expression of race-conserved *A. candida* CCG effectors to identify those that elicit HR when interacting with either of the broad-spectrum TNLs *WRR4A* and *WRR4B*. Both *R*-alleles were found to be capable of recognising multiple pathogenic CCGs, with eight detected by *WRR4A* and four additional CCGs by *WRR4B*. Of these established *Avr* elicitors, nearly all show some conservation between the three race 2 isolates Ac2v, AcBj12 and AcBjDC. Candidates such as CCG30 for example (recognised by *WRR4A*) are fully conserved between isolates here, as are CCG61 and CCG70 which are recognised by *WRR4B*. The latter CCG70 was found to provide an enhanced HR relative to other CCGs. Control of Donskaja resistance-breaking isolates Ac2v and AcBjDC (as well as AcBj12) would therefore be provided by interaction with either *WRR4A* or *WRR4B*. A study by, Castel *et al*. (2021 submitted) identified a *WRR4A* homologue in *Arabidopsis* ecotypes HR-5 and Oy-0 termed $WRR4A^{Oy\text{-}0}$, that contains a C-terminal extension which enables recognition of additional CCGs, including those of the *WRR4A*-virulent isolate AcEx1. Furthermore, *WRR8* and *WRR9* have been identified in *Arabidopsis* accessions Sf-2 and Hi-0 respectively, that confer complete resistance to Ac2v (Cevik *et al*., 2019).

A selection of *R*-alleles has therefore recently become established that each provide either complete or partial resistance against Donskaja-virulent race 2 isolates,

including *WRR4A*, *WRR4B*, *WRR4A*$^{Oy\text{-}0}$, *WRR8, WRR9* and the EH_25_DC.A06 *R*-allele (which requires confirmation). Combining all these together in a single gene stack would likely confer a significant fitness costs to the host plant (Burdon & Thrall, 2003). An optimal stack might therefore contain three *R*-genes; one from *Arabidopsis* - *WRR4A*$^{Oy\text{-}0}$ (others being redundant or known to be breakable), one from *B. juncea* – *BjuWRR1* (as the only one available) and one from *B. rapa* – the EH_25_DC.A06 *R*-allele. Provision of this GM stack alongside an *S*-gene knockout would add major enhancements to durability of resistance and may potentially provide too-great a hurdle for the pathogen to overcome. A GM approach could therefore provide great benefits to white rust resistance in Indian oilseed mustard, as well as other globally important *Brassica* crops.

Chapter 5

General Discussion

Accessing the rich genetic diversity contained within plant genetic resources (PGRs) is a prerequisite for successful breeding of key agricultural traits such as disease resistance. Modern advances in DNA sequencing technology and high-throughput marker analysis have become essential to this process, making it increasingly feasible to interrogate the genetic diversity contained within genebanks for mapping and identification of resistance genes. The wide availability of such genes is necessary for plant breeders to develop crop varieties containing durable resistance. These varieties will enhance sustainable intensification by alleviating the need for widespread application of environmentally damaging agrochemical treatments for disease control. Instead, a more targeted approach to pesticide application can be used alongside durable resistance to optimise protection of crops. In this way, the effectiveness of both genetic and chemical treatments can be preserved by alleviating selection pressure on pathogen adaption to overcome both forms of control.

The practice of introgressing single dominant or semi-dominant resistance genes into crop varieties has been undertaken throughout the latter half of the 20th century using conventional breeding practices. This frequently results in breakdown of the major gene resistance in so called 'boom-and-bust' cycles, where pathogen variants are selected for that have mutated or lost the corresponding *Avr* effector allele. The 'boom' can therefore be thought of as the widespread geographical distribution of the *R*-gene, typically throughout monoculture cropping systems. Whilst the 'bust' corresponds with a pathogen variant in the population overcoming this *R*-gene and rapidly propagating as the better adapted variant. This phenomenon is particularly well documented in cereal crops with regards to fungal rusts (Kolmer, 1996) and powdery mildews (Brown *et al*., 1996). Such boom-and-bust cycles are therefore counterproductive to resistance breeding, not only driving evolution of virulent pathogen isolates, but also leading to the permanent loss in commercial value of the broken *R*-gene.

Combining multiple genes into a single cultivar can enhance durability of disease control, though this approach is dependent on the availability of *R*-genes in a given breeding program and deployment of alleles within a crop (Mundt *et al*., 2002), as well as population structure and diversity of the pathogen (McDonald & Linde, 2002). In turn, *R*-allele availability relies to a large extent on the efficiency of trait mapping within large

genebank repositories and in particular within genetically diverse material such as landraces and wild species that may contain an abundance of currently 'untapped' alleles.

The availability of multiple sources of resistance genes within breeding programs provides opportunities for their strategic and targeted deployment over space and time. This is necessary as race specific *R*-gene function is dependent on the availability of the corresponding pathogen effector within local cropping systems. Emergence of virulent isolates is evidence therefore of how *R*-gene function can be overcome through functional mutations in corresponding effector genes. The relevance of this can be seen with the previous identification of broad-spectrum resistance gene *WRR4* $^{Col-0}$ in *A. thaliana* (Borhan *et al*., 2008), where three isolates where subsequently discovered that were able to overcome *WRR4* resistance (Fairhead, 2016). One of these isolates (AcEx1) was used to identify a new source of resistance in accession Oystese (Oy-0)(Fairhead, 2016), which has recently been identified as an alternative allele of WRR4A $^{Col-0}$ (termed *WRR4A*$^{Oy-0}$)(Castel *et al*., 2021 submitted). This TIR allele possesses a C-terminal extension that enables recognition of at least one additional CCG effector to that of WRR4A $^{Col-0}$, providing immunity to resistance breaking *A. candida* isolates.

This method of using identified resistance-breaking isolates is paramount to the targeted deployment of *R*-alleles by ensuring that multiple resistance specificities are available for breeding programs. Currently, the only molecularly characterised WRR gene in *B. juncea* is the CNL *BjuWRR1* from Donskaja, which is currently being considered for introgression into commercial varieties of Indian oilseed mustard. However, race 2 isolates have recently been found that break *BjuWRR1* resistance, which include the four Indian isolates Ac-BjBio-Pant, Ac-BjV-Pant, Ac-Bna-Pant and Ac-Alw (Dev *et al*., 2020), as well Canadian isolate Ac2v (Links *et al*., 2011) and the UK isolate AcBjDC. Here, mapping in *B. rapa* using AcBjDC identified the EH_25_DC.A06 locus in wild species accession EH_25 that contains a probable CNL WRR allele. This could have strategic importance for breeding durable white rust control in Indian *B. juncea* cultivars, if implemented alongside *BjuWRR1* within *B. juncea* varieties.

Future studies should therefore prioritise using the other mentioned *BjuWRR1*-virulent isolates for screening and mapping sources of WRR. It would be useful to know, for example, whether any of the resistance specificities identified here are shared with the other resistance breaking isolates. Mapping the same chromosome 6 locus using both

AcBj12 and AcBjDC highlights the feasibility of identifying a shared locus using the WGS-BSA method that provides potential control of different isolates. However, whether these *WRR* loci in *B. rapa* provide broad-spectrum resistance to other isolates remains unknown. The necessary resources exist in the UK to establish whether Ac2v resistance is found in any of the tested *B. rapa* accessions. Mapping using this isolate with the same material could potentially identify QTLs shared with those discovered here, adding further value to these loci as targets for *B. juncea* resistance breeding. Alternatively, seed from these *B. rapa* accessions could be sent to India for similar WGS-BSA testing with their *BjuWRR1*-virulent isolates.

Breeding strategies can implement the resistances identified in this study as they are, by pyramiding functional alleles from the multiple loci of *R-alleles* using conventional marker assisted breeding. Flanking markers, such as those denoted by the QTLseqr outputs (posPeakDeltaSNP and posMaxGprime) in Table 2.6, or those within candidate polymorphic *R*-genes could be used to achieve this. This strategy of marker assisted selection would require combining relevant *B. rapa* resistances within a single line for re-synthesis of *B. juncea*, followed by combining with Donskaja resistance for introgression into commercial cultivars. The alternative transgenic approach would require prior identification of the EH_25_DC.A06 *WRR* allele for inclusion within a single 'stacked' construct that includes *BjuWRR1*. Transformation of this construct into a crop variety would advance introgression of the trait by several years relative to conventional breeding, as well as eliminating the unwanted effects of linkage drag.

Advances in Next-Generation-Sequencing (NGS) technology have dramatically enhanced the process of mapping *R*-genes, where the affordability of producing vast SNP databases means that marker availability is no longer a limiting factor for trait mapping. The development of reduced-representation technologies such as GBS, alongside improved barcoding methods have come a long way to making NGS widely available, allowing multiplexing of numerous samples and bypassing issues of repetition and complexity within plant genomes. Today, resequencing of entire plant transcriptomes or whole genomes is becoming ever more feasible as prices continue to fall, maximising the availability of SNPs for analysis. Given the progression of the technology, it is reasonable to expect future whole-genome-resequencing (WGS) of complete genebank collections, providing insight from numerous complex plant genomes. Establishing these resources

within genebank documentation systems will add substantial value to collections, where combined datasets that include taxonomic, phenotypic and ecological data will provide a markedly improved breeding resource. For example, acquisition of such datasets will benefit novel breeding methods such as genomic selection by improving the genomic-estimated breeding value prediction accuracies required for successful breeding of complex traits (Bhat *et al*., 2016).

Computational and statistical advances in marker-trait analyses are becoming increasingly efficient at processing the deluge of genotype information made available by NGS tools. New methods such as QTL-seq have adapted NGS datasets with bulked segregant mapping (NGS-BSA) to allow rapid identification of QTLs from phenotypically variable resistant and susceptible tissue samples. The financial savings incurred from sampling only two tissue bulks per accession allow for screening of more accessions in parallel to maximise sampling of germplasm diversity for *R*-gene discovery. This study provided the first known application of QTL-seq methods (WGS-BSA) directly to genetically undeveloped genebank resources, which included a landrace (EH_5) and a wild species (EH_25). The identification in *Brassica rapa* of 16 high-resolution QTLs in this study, between five diverse genebank accessions represents an impressive mapping return within approximately a five-month experimental time frame. The result demonstrates the impressive potential of combining WGS-BSA methods with genebank resources to rapidly map novel resistance genes directly from undeveloped germplasm.

In this study, efforts to elucidate candidate genes within QTLs was limited by the exclusive use of tissue bulks, with markers applied to segregating individuals required for fine-mapping of individual candidate genes. This issue is compounded by the tendency of NLRs to cluster in the genome where NGS-BSA methods can only identify at the level of *R*-gene cluster rather than gene. It is recommended that future research adopts a hybrid mapping approach that includes both WGS-BSA and linkage methods. An optimal strategy for future mapping efforts is therefore described by the following steps: 1) Screen genebank resources for accessions that segregate for resistance; 2) Use a broadly susceptible accession to produce F_1 crosses with any discovered resistances to ascertain the F_1 phenotype and identify recessive resistances; 3) Prioritise recessive resistances (whilst simultaneously developing dominant lines where resources allow for production of tissue bulks using original parental phenotypic variation and generate WGS; 4) Advance

lines sent for sequencing for production of F_2 segregants; 5) Identify QTLs from WGS datasets and develop internal markers within polymorphic candidate genes; and 6) Screen markers back onto individual F_2 segregants to fine-map resistance.

From a breeding perspective, research efforts to identify causal genes may not be warranted, particularly where mapping has provided sufficiently closely linked markers that can be applied for MAS introgression of the trait. The use of WGS datasets in this study has produced virtually complete SNP data across all mapped QTLs, essentially providing access to the most tightly linked trait-markers for use in future MAS and breeding programs. There are obvious benefits to identifying a causal *R*-gene however, not only to provide an exact marker for MAS breeding or gene selection for GM stacking, but also to progress understanding of innate immunity mechanisms. The *B. oleracea* ACA2 resistance remains an example of this, where there is still potential to identify and characterise a novel mechanism of recessive white rust resistance.

The development of CRISPR genome editing technology has been pivotal in confirming the function of candidate genes. CRISPR has great potential for directly editing recessive susceptibility (*S*) genes in advanced cultivars, where loss-of-function can induce potentially durable sources of resistances. This represents a marked saving in time and resources relative to traditional breeding methods, even with use of MAS. For example, it would now be possible to induce the recessive race non-specific resistance found within EBH527 (ACA2) with the race specific resistance provided by *WRR4* $^{Col\text{-}0}$. The use of two different recognition specificities would provide a more durable resistance as the pathogen would be required to overcome control of both genes. Synthetic biology is also emerging as a promising tool for designing resistance alleles, whereby point mutations can be introduced that mimic naturally occurring resistances from wild species. Design of the synthetic *A. thaliana eIF4E1* allele for example was achieved using PCR-based mutagenesis to induce multiple amino acid changes that were based on resistance to potyvirus in naturally occurring *Pisum sativum* alleles (Bastet *et al*., 2018). The synthetic allele was able to extend the resistance spectrum to potyvirus isolates that lacked any previously established resistance, as well as to resistance-breaking isolates and an unrelated virus of the *Luteoviridae* family.

Monitoring pathogen diversity within any cropping system is an important prerequisite for anticipatory resistance breeding in response to emerging virulent

pathotypes. Third generation sequencing technologies such as Oxford Nanopore Technology platforms will play an increasingly pivotal role in achieving this, allowing rapid and affordable assembly of high-quality pathogen genomes from field isolates (Mongan *et al*., 2020). Comparative genomics can then be used alongside pathology and differential testing of isolates to identify mutated *Avr* elicitors that allow virulent pathotypes to avoid *R*-gene detection. The pathogenomics dataset in Chapter 4 is an example of this (Table 4.3), whereby future addition of Donskaja-avirulent effectoromes would advance shortlisting and discovery of the *BjuWRR1 Avr* elicitor. Results from this would aid the discovery of new *R*-genes that recognise this elicitor via transient expression *in planta*, as well as the decision-making process regarding optimal combinations of *R*-alleles for inclusion within any gene stack.

As discussed in Chapter 4, a suite of WRR alleles have recently been established (including *WRR4A*, *WRR4B*, *WRR4A*$^{Oy\text{-}0}$, *WRR8 and WRR9*) that each independently provide complete or near-complete protection against race 2 isolates (Borhan *et al*., 2008; Cevik *et al*., 2019; Redcar *et al*., unpublished 2021; Castel *et al*., 2021 submitted). This includes Ac2v, which we have shown here is a *BjuWRR1* resistance-breaking isolate (Figure 2.2). We have also shown that recognition specificities of these *R*-alleles include conserved CCG examples between the race 2 isolates examined in Chapter 4, such as CCG30, 61 or 70. An optimal gene stack for GM transformation of *B. juncea* varieties should therefore include one of these broad-spectrum *Arabidopsis R*-alleles, with *WRR4A*$^{Oy\text{-}0}$ suggested as the best suited of these due to enhanced recognition specificity.

In this study, methods have been developed that enhance the utility of genebank diversity for rapid mapping of resistance genes. This improves the availability of breeding resources required for targeted deployment of durable crop resistance against rapidly evolving plant pathogens. Resistances identified here have the potential to sustainably intensify production of Indian oilseed mustard, as well as other globally important *Brassica* oilseed and vegetable crops.

Bibliography

Aarts, N., Metz, M., Holub, E., Staskawicz, B. J., Daniels, M. J., & Parker, J. E. (1998). Different requirements for EDS1 and NDR1 by disease resistance genes define at least two R gene-mediated signaling pathways in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(17), 10306–10311. https://doi.org/10.1073/pnas.95.17.10306

Adachi, H., Derevnina, L., & Kamoun, S. (2019). NLR singletons, pairs, and networks: evolution, assembly, and regulation of the intracellular immunoreceptor circuitry of plants. *Current Opinion in Plant Biology*, *50*, 121–131. https://doi.org/10.1016/j.pbi.2019.04.007

Adhikari, T. B., Liu, J. Q., Mathur, S., Wu, C. X., & Rimmer, S. R. (2003). Genetic and molecular analyses in crosses of race 2 and race 7 of Albugo candida. *Phytopathology*, *93*(8), 959–965. https://doi.org/10.1094/PHYTO.2003.93.8.959

Afanador, L., Haley, S., & Kelly, J. D. (1993). Adoption of a "mini-prep" DNA extraction method for RAPD marker analysis in Common Bean (Phaseolus vulgar is L.). *Annual Report of the Bean Improvement Cooperative*, *6*(36), 10–11.

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, *37*(4), 420–423. https://doi.org/10.1038/s41587-019-0036-z

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. https://doi.org/10.1038/nrg.2015.28

Anjum, N. A., Gill, S. S., Ahmad, I., Pacheco, M., Duarte, A. C., Umar, S., Khan, N. A., & Pereira, M. E. (2012). *The Plant Family Brassicaceae: An Introduction*. https://doi.org/10.1007/978-94-007-3913-0_1

Arora, H., Padmaja, K. L., Paritosh, K., Mukhi, N., Tewari, A. K., Mukhopadhyay, A., Gupta, V., Pradhan, A. K., & Pental, D. (2019). BjuWRR1, a CC-NB-LRR gene identified in *Brassica juncea*, confers resistance to white rust caused by Albugo candida. *Theoretical and Applied Genetics*, *132*(8), 2223–2236. https://doi.org/10.1007/s00122-019-03350-z

Arya, P., Kumar, G., Acharya, V., & Singh, A. K. (2014). Genome-wide identification and expression analysis of nbs-encoding genes in malus X domestica and expansion of nbs genes family in rosaceae. *PLoS ONE*, *9*(9), 502. https://doi.org/10.1371/journal.pone.0107987

Baggs, E., Dagdas, G., & Krasileva, K. V. (2017). NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Current Opinion in Plant Biology*, *38*(Figure 1), 59–67. https://doi.org/10.1016/j.pbi.2017.04.012

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, *37*(SUPPL. 2), 202–208. https://doi.org/10.1093/nar/gkp335

Bailey, T. L., & Gribskov, M. (1998). Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics*, *14*(1), 48–54. https://doi.org/10.1093/bioinformatics/14.1.48

Balint-Kurti, P. (2019). The plant hypersensitive response: concepts, control and consequences. *Molecular Plant Pathology*, *20*(8), 1163–1178. https://doi.org/10.1111/mpp.12821

Barre, A., Bourne, Y., Van Damme, E. J. M., & Rougé, P. (2019). Overview of the structure–Function relationships of mannose-specific lectins from plants, Algae and Fungi. *International Journal of Molecular Sciences*, *20*(2). https://doi.org/10.3390/ijms20020254

Bastet, A., Lederer, B., Giovinazzo, N., Arnoux, X., German-Retana, S., Reinbold, C., Brault, V., Garcia, D., Djennane, S., Gersch, S., Lemaire, O., Robaglia, C., & Gallois, J. L. (2018). Trans-species synthetic gene design allows resistance pyramiding and broad-spectrum engineering of virus resistance in plants. *Plant Biotechnology Journal*, *16*(9), 1569–1581. https://doi.org/10.1111/pbi.12896

Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., & Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis* thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(43), 18724–18728. https://doi.org/10.1073/pnas.0909766107

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P., & Prabhu, K. V. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Frontiers in Genetics*, *7*(DEC), 1–11. https://doi.org/10.3389/fgene.2016.00221

Blanvillain-Baufumé, S., Reschke, M., Solé, M., Auguy, F., Doucoure, H., Szurek, B., Meynard, D., Portefaix, M., Cunnac, S., Guiderdoni, E., Boch, J., & Koebnik, R. (2017). Targeted promoter editing for rice resistance to Xanthomonas oryzae pv. oryzae reveals differential activities for SWEET14-inducing TAL effectors. *Plant Biotechnology Journal*, *15*(3), 306–317. https://doi.org/10.1111/pbi.12613

Bolton, M. D., de Jonge, R., Inderbitzin, P., Liu, Z., Birla, K., Van de Peer, Y., Subbarao, K. V., Thomma, B. P. H. J., & Secor, G. A. (2014). The heterothallic sugarbeet pathogen Cercospora beticola contains exon fragments of both MAT genes that are homogenized by concerted evolution. *Fungal Genetics and Biology*, *62*, 43–54. https://doi.org/10.1016/j.fgb.2013.10.011

Borhan, M. H., Gunn, N., Cooper, A., Gulden, S., Tör, M., Rimmer, S. R., & Holub, E. B. (2008). WRR4 encodes a TIR-NB-LRR protein that confers broad-spectrum white rust resistance in *Arabidopsis thaliana* to four physiological races of Albugo candida. *Molecular*

Plant-Microbe Interactions, *21*(6), 757–768. https://doi.org/10.1094/MPMI-21-6-0757

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Branham, S. E., & Farnham, M. W. (2019). Identification of heat tolerance loci in broccoli through bulked segregant analysis using whole genome resequencing. *Euphytica*, *215*(2), 1–9. https://doi.org/10.1007/s10681-018-2334-9

Branham, S. E., Patrick Wechter, W., Lambel, S., Massey, L., Ma, M., Fauve, J., Farnham, M. W., & Levi, A. (2018). QTL-seq and marker development for resistance to Fusarium oxysporum f. sp. niveum race 1 in cultivated watermelon. *Molecular Breeding*, *38*(11). https://doi.org/10.1007/s11032-018-0896-9

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1842*(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

Burdon, J. J., & Thrall, P. H. (2003). The fitness costs to plants of resistance to pathogens. *Genome Biology*, *4*(9). https://doi.org/10.1186/gb-2003-4-9-227

Büschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., Van Daelen, R., Van der Lee, T., Diergaarde, P., Groenendijk, J., Töpsch, S., Vos, P., Salamini, F., & Schulze-Lefert, P. (1997). The barley Mlo gene: A novel control element of plant pathogen resistance. *Cell*, *88*(5), 695–705. https://doi.org/10.1016/S0092-8674(00)81912-1

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *2018*(6). https://doi.org/10.7717/peerj.4958

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., & May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biology*, *4*, 1–21. https://doi.org/10.1186/1471-2229-4-10

Cantor, R. M. (2018). Analysis of genetic linkage. In *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics: Foundations* (Seventh Ed). Elsevier Inc. https://doi.org/10.1016/B978-0-12-812537-3.00008-1

Cevik, V., Boutrot, F., Apel, W., Robert-Seilaniantz, A., Furzer, O. J., Redkar, A., Castel, B., Kover, P. X., Prince, D. C., Holub, E. B., & Jones, J. D. G. (2019). Transgressive segregation reveals mechanisms of *Arabidopsis* immunity to *Brassica*-infecting races of white rust (Albugo candida). *Proceedings of the National Academy of Sciences of the United States of America*, *116*(7), 2767–2773. https://doi.org/10.1073/pnas.1812911116

Chandrasekaran, J., Brumin, M., Wolf, D., Leibman, D., Klap, C., Pearlsman, M., Sherman, A., Arazi, T., & Gal-On, A. (2016). Development of broad virus resistance in non-transgenic

cucumber using CRISPR/Cas9 technology. *Molecular Plant Pathology*, *17*(7), 1140–1153. https://doi.org/10.1111/mpp.12375

Chen, Z., Erickson, D. L., & Meng, J. (2020). Benchmarking long-read assemblers for genomic analyses of bacterial pathogens using oxford nanopore sequencing. *International Journal of Molecular Sciences*, *21*(23), 1–27. https://doi.org/10.3390/ijms21239161

Cheng, F., Wu, J., & Wang, X. (2014). Genome triplication drove the diversification of *Brassica* plants. *Horticulture Research*, *1*(February), 1–8. https://doi.org/10.1038/hortres.2014.24

Chisholm, S. T., Coaker, G., Day, B., & Staskawicz, B. J. (2006). Host-microbe interactions: Shaping the evolution of the plant immune response. *Cell*, *124*(4), 803–814. https://doi.org/10.1016/j.cell.2006.02.008

Choi, Y.-J., Shin, H.-D., Ploch, S., & Thines, M. (2008). Evidence for uncharted biodiversity in the Albugo candida complex, with the description of a new species. *Mycological Research*, *112*(11), 1327–1334. https://doi.org/10.1016/j.mycres.2008.04.015

Cohen, J. I., & Loskutov, I. G. (2016). Exploring the nature of science through courage and purpose: a case study of Nikolai Vavilov and plant biodiversity. *SpringerPlus*, *5*(1). https://doi.org/10.1186/s40064-016-2795-z

Cooper, A. J., Latunde-Dada, A. O., Woods-Tör, A., Lynn, J., Lucas, J. A., Crute, I. R., & Holub, E. B. (2008). Basic compatibility of *Albugo candida* in *Arabidopsis thaliana* and *Brassica juncea* causes broad-spectrum suppression of innate immunity. *Molecular Plant-Microbe Interactions*, *21*(6), 745–756. https://doi.org/10.1094/MPMI-21-6-0745

Copping, L. (1998). Systemic acquired resistance. *Pesticide Outlook*, *9*(2), 34. https://doi.org/10.2307/3870231

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dangl, J. L., & Jones, J. D. G. (2001). Defence Responses To Infection. *Nature*, *411*(June).

Datema, E., Hulzink, R., Blommers, L., Espejo Valle-Inclan, J., van Orsouw, N., Wittenberg, A., & de Vos, M. (2016). The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. *BioRxiv*, 084772. https://doi.org/10.1101/084772

Dev, D., Tewari, A. K., Upadhyay, P., & Daniel, G. R. (2020). Identification and nomenclature of Albugo candida pathotypes of Indian origin causing white rust disease of rapeseed-mustard. *European Journal of Plant Pathology*, *158*(4), 987–1004. https://doi.org/10.1007/s10658-020-02135-1

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), 1–10. https://doi.org/10.1371/journal.pone.0019379

Engler, C., Kandzia, R., & Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE*, *3*(11). https://doi.org/10.1371/journal.pone.0003647

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2), 29–37. https://doi.org/10.1093/nar/gkr367

Fitt, B. D. L., Brun, H., Barbetti, M. J., & Rimmer, S. R. (2006). World-wide importance of phoma stem canker (*Leptosphaeria maculans and L. biglobosa*) on oilseed Rape (*Brassica napus*). *European Journal of Plant Pathology*, *114*(1), 3–15. https://doi.org/10.1007/s10658-005-2233-5

Flor, H. H. (1971). *Current status of the gene-fob-gene concept*. 275–296.

Frankel, O. H., & Brown, A. H. D. (1984). Current plant genetic resources - a critical appraisal. *Proceedings XV International Congress of Genetics*, *4*(September), 1–13.

Gairola, K., & Tewari, A. K. (2019). Management of white rust (Albugo candida) in indian mustard by fungicides and garlic extract. *Pesticide Research Journal*, *31*(1), 60–65. https://doi.org/10.5958/2249-524X.2019.00006.2

Gao, M., Yin, X., Yang, W., Lam, S. M., Tong, X., Liu, J., Wang, X., Li, Q., Shui, G., & He, Z. (2017). GDSL lipases modulate immunity through lipid homeostasis in rice. *PLoS Pathogens*, *13*(11), 1–24. https://doi.org/10.1371/journal.ppat.1006724

Geu-Flores, F., Nour-Eldin, H. H., Nielsen, M. T., & Halkier, B. A. (2007). USER fusion: A rapid and efficient method for simultaneous fusion and cloning of multiple PCR products. *Nucleic Acids Research*, *35*(7), 0–5. https://doi.org/10.1093/nar/gkm106

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), 1178–1186. https://doi.org/10.1093/nar/gkr944

Guo, Y. L., Fitz, J., Schneeberger, K., Ossowski, S., Cao, J., & Weigel, D. (2011). Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*. *Plant Physiology*, *157*(2), 757–769. https://doi.org/10.1104/pp.111.181990

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool

for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*(18), 4606–4618. https://doi.org/10.1111/mec.12415

Hochberg, Y. (2016). *Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 (1995), Publi*. *57*(1), 289–300.

Holub, E. B., Brose, E., Tor, M., Clay, C., Crute, I. R., & Beynon, J. L. (1995). Phenotypic and genotypic variation in the interaction between *Arabidopsis thaliana* and *Albugo candida*. In *Molecular Plant-Microbe Interactions* (Vol. 8, Issue 6, pp. 916–928). https://doi.org/10.1094/MPMI-8-0916

Hummer, K. E. (2015). In the footsteps of vavilov: Plant diversity then and now. *HortScience*, *50*(6), 784–788. https://doi.org/10.21273/hortsci.50.6.784

Hwang, I. S., & Hwang, B. K. (2011). The pepper mannose-binding lectin gene CaMBL1 is required to regulate cell death and defense responses to microbial pathogens. *Plant Physiology*, *155*(1), 447–463. https://doi.org/10.1104/pp.110.164848

Illa-Berenguer, E., Van Houten, J., Huang, Z., & van der Knaap, E. (2015). Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. *Theoretical and Applied Genetics*, *128*(7), 1329–1342. https://doi.org/10.1007/s00122-015-2509-x

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 1–11. https://doi.org/10.1186/s13059-016-1103-0

Jia, H., Orbovic, V., Jones, J. B., & Wang, N. (2016). Modification of the PthA4 effector binding elements in Type I CsLOB1 promoter using Cas9/sgRNA to produce transgenic Duncan grapefruit alleviating XccΔpthA4: DCsLOB1.3 infection. *Plant Biotechnology Journal*, *14*(5), 1291–1301. https://doi.org/10.1111/pbi.12495

Jia, H., Zhang, Y., Orbović, V., Xu, J., White, F. F., Jones, J. B., & Wang, N. (2017). Genome editing of the disease susceptibility gene CsLOB1 in citrus confers resistance to citrus canker. *Plant Biotechnology Journal*, *15*(7), 817–823. https://doi.org/10.1111/pbi.12677

Jiang, W., Zhou, H., Bi, H., Fromm, M., Yang, B., & Weeks, D. P. (2013). Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in *Arabidopsis*, tobacco, sorghum and rice. *Nucleic Acids Research*, *41*(20), 1–12. https://doi.org/10.1093/nar/gkt780

Jones, J.D.G., D. J. L. (2006). H E P L a N T Imm U N E S Y S T. *Nature*, *444*, 323–329.
Jørgensen, I. H. (1992). Discovery, characterization and exploitation of Mlo powdery mildew resistance in barley. *Euphytica*, *63*(1–2), 141–152. https://doi.org/10.1007/BF00023919

Jouet, A., Saunders, D. G. O., McMullan, M., Ward, B., Furzer, O., Jupe, F., Cevik, V., Hein, I., Thilliez, G. J. A., Holub, E., van Oosterhout, C., & Jones, J. D. G. (2019). Albugo candida race diversity, ploidy and host-associated microbes revealed using DNA sequence capture on diseased plants in the field. *New Phytologist*, *221*(3), 1529–1543. https://doi.org/10.1111/nph.15417

Jupe, F., Witek, K., Verweij, W., Śliwka, J., Pritchard, L., Etherington, G. J., Maclean, D., Cock, P. J., Leggett, R. M., Bryan, G. J., Cardle, L., Hein, I., & Jones, J. D. G. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant Journal*, *76*(3), 530–544. https://doi.org/10.1111/tpj.12307

Kalia, P., Saha, P., & Ray, S. (2017). Development of RAPD and ISSR derived SCAR markers linked to Xca1Bo gene conferring resistance to black rot disease in cauliflower (*Brassica oleracea var. botrytis L.*). *Euphytica*, *213*(10), 1–12. https://doi.org/10.1007/s10681-017-2025-y

Kaminski, K. P., Kørup, K., Andersen, M. N., Sønderkær, M., Andersen, M. S., Kirk, H. G., & Nielsen, K. L. (2015). Cytosolic glutamine synthetase is important for photosynthetic efficiency and water use efficiency in potato as revealed by high-throughput sequencing QTL analysis. *Theoretical and Applied Genetics*, *128*(11), 2143–2153. https://doi.org/10.1007/s00122-015-2573-2

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kemen, E., Gardiner, A., Schultz-Larsen, T., Kemen, A. C., Balmuth, A. L., Robert-Seilaniantz, A., Bailey, K., Holub, E., Studholme, D. J., MacLean, D., & Jones, J. D. G. (2011). Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biology*, *9*(7). https://doi.org/10.1371/journal.pbio.1001094

Kilian, B., & Graner, A. (2012). NGS technologies for analyzing germplasm diversity in genebanks. *Briefings in Functional Genomics*, *11*(1), 38–50.

https://doi.org/10.1093/bfgp/elr046

Kim, J., Kang, W. H., Hwang, J., Yang, H. B., Dosun, K., Oh, C. S., & Kang, B. C. (2014). Transgenic *Brassica rapa* plants over-expressing eIF(iso)4E variants show broad-spectrum Turnip mosaic virus (TuMV) resistance. *Molecular Plant Pathology*, *15*(6), 615–626. https://doi.org/10.1111/mpp.12120

Kole, C., Thormann, C. E., Karlsson, B. H., Palta, J. P., Gaffney, P., Yandell, B., & Osborn, T. C. (2002). Comparative mapping of loci controlling winter survival and related traits in oilseed *Brassica rap*a and *B. napus*. *Molecular Breeding*, *9*(3), 201–210. https://doi.org/10.1023/A:1019759512347

Kwon, S. J., Jin, H. C., Lee, S., Nam, M. H., Chung, J. H., Kwon, S. Il, Ryu, C. M., & Park, O. K. (2009). GDSL lipase-like 1 regulates systemic resistance associated with ethylene signaling in *Arabidopsis*. *Plant Journal*, *58*(2), 235–245. https://doi.org/10.1111/j.1365-313X.2008.03772.x

Lai, Y., Dang, F., Lin, J., Yu, L., Shi, Y., Xiao, Y., Huang, M., Lin, J., Chen, C., Qi, A., Liu, Z., Guan, D., Mou, S., Qiu, A., & He, S. (2013). Overexpression of a Chinese cabbage BrERF11 transcription factor enhances disease resistance to Ralstonia solanacearum in tobacco. *Plant Physiology and Biochemistry*, *62*, 70–78. https://doi.org/10.1016/j.plaphy.2012.10.010

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. https://doi.org/10.1016/j.bdq.2015.02.001

Lawrenson, T., Shorinola, O., Stacey, N., Li, C., Østergaard, L., Patron, N., Uauy, C., & Harwood, W. (2015). Induction of targeted, heritable mutations in barley and *Brassica* oleracea using RNA-guided Cas9 nuclease. *Genome Biology*, *16*(1), 1–13. https://doi.org/10.1186/s13059-015-0826-7

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Links, M. G., Holub, E., Jiang, R. H. Y., Sharpe, A. G., Hegedus, D., Beynon, E., Sillito, D., Clarke, W. E., Uzuhashi, S., & Borhan, M. H. (2011). De novo sequence assembly of Albugo candida reveals a small genome relative to other biotrophic oomycetes. *BMC Genomics*, *12*(1), 503. https://doi.org/10.1186/1471-2164-12-503

Liu, Y., Xu, A., Liang, F., Yao, X., Wang, Y., Liu, X., Zhang, Y., Dalelhan, J., Zhang, B., Qin, M., Huang, Z., & Shaolin, L. (2018). Screening of clubroot-resistant varieties and transfer of clubroot resistance genes to *Brassica napus* using distant hybridization. *Breeding Science*, *68*(2), 258–267. https://doi.org/10.1270/jsbbs.17125

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., Sun, J., Zhang, Z., Weng, Y., & Huang, S. (2014). QTL-seq identifies an early flowering QTL located near flowering locus T in cucumber. *Theoretical and Applied Genetics*, *127*(7), 1491–1499. https://doi.org/10.1007/s00122-014-2313-z

LUO, X. dong, LIU, J., ZHAO, J., DAI, L. fang, CHEN, Y. ling, ZHANG, L., ZHANG, F., HU, B. lin, & XIE, J. kun. (2018). Rapid mapping of candidate genes for cold tolerance in Oryza rufipogon Griff. by QTL-seq of seedlings. In *Journal of Integrative Agriculture* (Vol. 17, Issue 2, pp. 265–275). https://doi.org/10.1016/S2095-3119(17)61712-X

Lv, H., Fang, Z., Yang, L., Zhang, Y., Wang, Q., Liu, Y., Zhuang, M., Yang, Y., Xie, B., Liu, B., Liu, J., Kang, J., & Wang, X. (2014). Mapping and analysis of a novel candidate Fusarium wilt resistance gene in. *BMC Genomics*, *15*(1), 1–10. https://doi.org/10.1186/1471-2164-15-1094

Lv, H., Fang, Z., Yang, L., Zhang, Y., & Wang, Y. (2020). An update on the arsenal: mining resistance genes for disease management of *Brassica* crops in the genomic era. *Horticulture Research*, *7*(1). https://doi.org/10.1038/s41438-020-0257-9

Lysak, M. A., Koch, M. A., Pecinka, A., & Schubert, I. (2005). Chromosome triplication found across the tribe Brassiceae. *Genome Research*, *15*(4), 516–525. https://doi.org/10.1101/gr.3531105

Magwene, P. M., Willis, J. H., & Kelly, J. K. (2011). The statistics of bulk segregant analysis using next generation sequencing. *PLoS Computational Biology*, *7*(11), 1–9. https://doi.org/10.1371/journal.pcbi.1002255

Malnoy, M., Viola, R., Jung, M. H., Koo, O. J., Kim, S., Kim, J. S., Velasco, R., & Kanchiswamy, C. N. (2016). DNA-free genetically edited grapevine and apple protoplast using CRISPR/Cas9 ribonucleoproteins. *Frontiers in Plant Science*, *7*(DECEMBER2016), 1–9. https://doi.org/10.3389/fpls.2016.01904

Mansfeld, B. N., & Grumet, R. (2018). QTLseqr: An R package for bulk segregant analysis with next-generation sequencing. *Plant Genome*, *11*(2), 1–5. https://doi.org/10.3835/plantgenome2018.01.0006

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), 1–10. https://doi.org/10.1002/mpr.1608

McHale, L., Tan, X., Koehl, P., & Michelmore, R. W. (2006). Plant NBS-LRR proteins: Adaptable guards. *Genome Biology*, *7*(4). https://doi.org/10.1186/gb-2006-7-4-212

McMullan, M., Gardiner, A., Bailey, K., Kemen, E., Ward, B. J., Cevik, V., Robert-Seilaniantz, A., Schultz-Larsen, T., Balmuth, A., Holub, E., van Oosterhout, C., & Jones, J. D. G. (2015). Evidence for suppression of immunity as a driver for genomic introgressions and host

range expansion in races of Albugo candida, a generalist parasite. *ELife*, *2015*(4), 1–24. https://doi.org/10.7554/eLife.04550

Mei, J., Liu, Y., Wei, D., Wittkop, B., Ding, Y., Li, Q., Li, J., Wan, H., Li, Z., Ge, X., Frauen, M., Snowdon, R. J., Qian, W., & Friedt, W. (2015). Transfer of sclerotinia resistance from wild relative of *Brassica oleracea* into *Brassica napus* using a hexaploidy step. *Theoretical and Applied Genetics*, *128*(4), 639–644. https://doi.org/10.1007/s00122-015-2459-3

Mendel, G. (1901). E p h (1865). *Scholarly Publishing*, *1865*, 3–47.

Meyers, B. C., Kozik, A., Griego, A., Kuang, H., & Michelmore, R. W. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell*, *15*(4), 809–834. https://doi.org/10.1105/tpc.009308

Meyers, B. C., Morgante, M., & Michelmore, R. W. (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 Journal*, *32*(1), 77–92. https://doi.org/10.1046/j.1365-313X.2002.01404.x

Michelmore, R. W., Paran, I., & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(21), 9828–9832. https://doi.org/10.1073/pnas.88.21.9828

Mongan, A. E., Tuda, J. S. B., & Runtuwene, L. R. (2020). Portable sequencer in the fight against infectious disease. *Journal of Human Genetics*, *65*(1), 35–40. https://doi.org/10.1038/s10038-019-0675-4

Müller, M., & Munné-Bosch, S. (2015). Ethylene response factors: A key regulatory hub in hormone and stress signaling. *Plant Physiology*, *169*(1), 32–41. https://doi.org/10.1104/pp.15.00677

Mundt, C. C., Cowger, C., & Garrett, K. A. (2002). Relevance of integrated disease management to resistance durability. *Euphytica*, *124*(2), 245–252. https://doi.org/10.1023/A:1015642819151

Nanjundan, J., Radhamani, J., Thakur, A. K., Berliner, J., Manjunatha, C., Sindhu, A., Aravind, J., & Singh, K. H. (2020). Utilization of Rapeseed-Mustard Genetic Resources for *Brassica* Improvement: A Retrospective Approach. In *Brassica Improvement*. https://doi.org/10.1007/978-3-030-34694-2_1

Nekrasov, V., Wang, C., Win, J., Lanz, C., Weigel, D., & Kamoun, S. (2017). Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion. *Scientific Reports*, *7*(1), 1–6. https://doi.org/10.1038/s41598-017-00578-x

Ntoukakis, V., Saur, I. M. L., Conlan, B., & Rathjen, J. P. (2014). The changing of the guard:

The Pto/Prf receptor complex of tomato and pathogen recognition. *Current Opinion in Plant Biology*, *20*, 69–74. https://doi.org/10.1016/j.pbi.2014.04.002

Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. L. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theoretical and Applied Genetics*, *126*(2), 289–305. https://doi.org/10.1007/s00122-012-1971-y

On, C., Resources, G., & Food, F. O. R. (2019). The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture. In *The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture*. https://doi.org/10.4060/i4787e

Ortigosa, A., Gimenez-Ibanez, S., Leonhardt, N., & Solano, R. (2019). Design of a bacterial speck resistant tomato by CRISPR/Cas9-mediated editing of SlJAZ2. *Plant Biotechnology Journal*, *17*(3), 665–673. https://doi.org/10.1111/pbi.13006

Osuna-Cruz, C. M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Cigliano, R. A., Sanseverino, W., & Ercolano, M. R. (2018). PRGdb 3.0: A comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Research*, *46*(D1), D1197–D1201. https://doi.org/10.1093/nar/gkx1119

Pandey, M. K., Khan, A. W., Singh, V. K., Vishwakarma, M. K., Shasidhar, Y., Kumar, V., Garg, V., Bhat, R. S., Chitikineni, A., Janila, P., Guo, B., & Varshney, R. K. (2017). QTL-seq approach identified genomic regions and diagnostic markers for rust and late leaf spot resistance in groundnut (Arachis hypogaea L.). *Plant Biotechnology Journal*, *15*(8), 927–941. https://doi.org/10.1111/pbi.12686

Panjabi-Massand, P., Yadava, S. K., Sharma, P., Kaur, A., Kumar, A., Arumugam, N., Sodhi, Y. S., Mukhopadhyay, A., Gupta, V., Pradhan, A. K., & Pental, D. (2010). Molecular mapping reveals two independent loci conferring resistance to Albugo Candida in the East European germplasm of oilseed mustard *Brassica juncea*. *Theoretical and Applied Genetics*, *121*(1), 137–145. https://doi.org/10.1007/s00122-010-1297-6

Parkin, I. A. P., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., Denoeud, F., Belcram, H., Links, M. G., Just, J., Clarke, C., Bender, T., Huebert, T., … Sharpe, A. G. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biology*, *15*(6), 1–18. https://doi.org/10.1186/gb-2014-15-6-r77

Paula de Toledo Thomazella, D., Brail, Q., Dahlbeck, D., & Staskawicz, B. (2016). *CRISPR-Cas9 mediated mutagenesis of a DMR6 ortholog in tomato confers broad-spectrum disease resistance*. 064824. https://doi.org/10.1101/064824

Peng, A., Chen, S., Lei, T., Xu, L., He, Y., Wu, L., Yao, L., & Zou, X. (2017). Engineering canker-resistant plants through CRISPR/Cas9-targeted editing of the susceptibility gene CsLOB1 promoter in citrus. *Plant Biotechnology Journal*, *15*(12), 1509–1519.

https://doi.org/10.1111/pbi.12733

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same. *National Review of Genetics*, *12*(1), 32–42. https://doi.org/10.1038/nrg2899.Synonymous

Poczai, P., Cernák, I., Varga, I., & Hyvönen, J. (2014). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Genetic Resources and Crop Evolution*, *61*(1), 796–815. https://doi.org/10.1134/S1022795411020074

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., … Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 1–22. https://doi.org/10.1101/201178

Porter, B. W., Paidi, M., Ming, R., Alam, M., Nishijima, W. T., & Zhu, Y. J. (2009). Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family. *Molecular Genetics and Genomics*, *281*(6), 609–626. https://doi.org/10.1007/s00438-009-0434-x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Pyott, D. E., Sheehan, E., & Molnar, A. (2016). Engineering of CRISPR/Cas9-mediated potyvirus resistance in transgene-free *Arabidopsis* plants. *Molecular Plant Pathology*, *17*(8), 1276–1288. https://doi.org/10.1111/mpp.12417

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Ramos, A., Fu, Y., Michael, V., & Meru, G. (2020). QTL-seq for identification of loci associated with resistance to Phytophthora crown rot in squash. *Scientific Reports*, *10*(1), 1–8. https://doi.org/10.1038/s41598-020-62228-z

Rédei, G. P. (2008). Central Dogma. *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, *227*, 309–309. https://doi.org/10.1007/978-1-4020-6754-9_2672

Reyes-Valdés, M. H., Burgueño, J., Singh, S., Martínez, O., & Sansaloni, C. P. (2018). An informational view of accession rarity and allele specificity in germplasm banks for management and conservation. *PLoS ONE*, *13*(2), 1–15. https://doi.org/10.1371/journal.pone.0193346

Rimmer, S. R., Mathur, S., & Wu, C. R. (2000). Virulence of isolates of Albugo candida from western Canada to *Brassica* species. *Canadian Journal of Plant Pathology*, *22*(3), 229–235. https://doi.org/10.1080/07060660009500468

Ronald, P. C., Salmeron, J. M., Carland, F. M., & Staskawicz, B. J. (1992). The cloned avirulence gene avrPto induces disease resistance in tomato cultivars containing the Pto resistance gene. *Journal of Bacteriology*, *174*(5), 1604–1611. https://doi.org/10.1128/jb.174.5.1604-1611.1992

Saharan, G. S., Verma, P. R., Meena, P. D., Borhan, M. H., & Singh, D. (2014). *Analysis of White Rust Research Progress Through Bibliography*. *5*(August 2016), 42–115.

Salmela, L., & Rivals, E. (2014). LoRDEC: Accurate and efficient long read error correction. *Bioinformatics*, *30*(24), 3506–3514. https://doi.org/10.1093/bioinformatics/btu538

Sandhya, D., Jogam, P., Allini, V. R., Abbagani, S., & Alok, A. (2020). The present and potential future methods for delivering CRISPR/Cas9 components in plants. *Journal of Genetic Engineering and Biotechnology*, *18*(1). https://doi.org/10.1186/s43141-020-00036-8

Santos, M. R., & Dias, J. S. (2004). Evaluation of a core collection of *Brassica* oleracea accessions for resistance to white rust of crucifers (Albugo candida) at the cotyledon stage. *Genetic Resources and Crop Evolution*, *51*(7), 713–722. https://doi.org/10.1023/B:GRES.0000034577.82868.5d

Santos, M. R., Dias, J. S., Silva, M. J., & Ferreira-Pinto, M. M. (2006). Resistance to white rust in pak choi and Chinese cabbage at the cotyledon stage. *Communications in Agricultural and Applied Biological Sciences*, *71*(3 Pt B), 963–971.

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, *12*(10), 683–691. https://doi.org/10.1038/nrg3051

Shan, W., Cao, M., Leung, D., & Tyler, B. M. (2004). The Avr1b locus of Phytophthora sojae encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene Rps1b. *Molecular Plant-Microbe Interactions*, *17*(4), 394–403. https://doi.org/10.1094/MPMI.2004.17.4.394

Sharma, S., Upadhyaya, H. D., Varshney, R. K., & Gowda, C. L. L. (2013). Pre-breeding for diversification of primary gene pool and genetic enhancement of grain legumes. *Frontiers in Plant Science*, *4*(AUG). https://doi.org/10.3389/fpls.2013.00309

Shekhawat, K., Rathore, S. S., Premi, O. P., Kandpal, B. K., & Chauhan, J. S. (2012). Advances in Agronomic Management of Indian Mustard (*Brassica juncea (L.) Czernj. Cosson*): An Overview . *International Journal of Agronomy*, *2012*, 1–14. https://doi.org/10.1155/2012/408284

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

https://doi.org/10.1093/bioinformatics/btv351

Smirnova, O. G., & Kochetov, A. V. (2015). Promoters of plant genes responsive to pathogen invasion. *Russian Journal of Genetics: Applied Research*, *5*(3), 254–261. https://doi.org/10.1134/S2079059715030181

Son, G. H., Wan, J., Kim, H. J., Nguyen, X. C., Chung, W. S., Hong, J. C., & Stacey, G. (2012). Ethylene-responsive element-binding factor 5, ERF5, is involved in chitin-induced innate immunity response. *Molecular Plant-Microbe Interactions*, *25*(1), 48–60. https://doi.org/10.1094/MPMI-06-11-0165

Spielmeyer, W., Sharp, P. J., & Lagudah, E. S. (2003). Identification and validation of markers linked to broad-spectrum stem rust resistance gene Sr2 in wheat (Triticum aestivum L.). *Crop Science*, *43*(1), 333–336. https://doi.org/10.2135/cropsci2003.3330

Srivastava, R., Upadhyaya, H. D., Kumar, R., Daware, A., Basu, U., Shimray, P. W., Tripathi, S., Bharadwaj, C., Tyagi, A. K., & Parida, S. K. (2017). A multiple QTL-Seq strategy delineates potential genomic loci governing flowering time in chickpea. *Frontiers in Plant Science*, *8*(July), 1–13. https://doi.org/10.3389/fpls.2017.01105

Staal, J., Kaliff, M., Dewaele, E., Persson, M., & Dixelius, C. (2008). RLM3, a TIR domain encoding gene involved in broad-range immunity of *Arabidopsis* to necrotrophic fungal pathogens. *Plant Journal*, *55*(2), 188–200. https://doi.org/10.1111/j.1365-313X.2008.03503.x

Staskawicz, B. J., Dahlbeck, D., & Keen, N. T. (1984). Cloned avirulence gene of Pseudomonas syringae pv. glycinea determines race-specific incompatibility on Glycine max (L.) Merr. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(19 I), 6024–6028. https://doi.org/10.1073/pnas.81.19.6024

Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G., & Wulff, B. B. H. (2015). NLR-parser: Rapid annotation of plant NLR complements. *Bioinformatics*, *31*(10), 1665–1667. https://doi.org/10.1093/bioinformatics/btv005

Swiderski, M. R., & Innes, R. W. (2001). The *Arabidopsis* PBS1 resistance gene encodes a member of a novel protein kinase subfamily. *Plant Journal*, *26*(1), 101–112. https://doi.org/10.1046/j.1365-313X.2001.01014.x

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., Uemura, A., Utsushi, H., Tamiru, M., Takuno, S., Innan, H., Cano, L. M., Kamoun, S., & Terauchi, R. (2013). QTL-seq: Rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant Journal*, *74*(1), 174–183. https://doi.org/10.1111/tpj.12105

Tang, X., Frederick, R. D., Zhou, J., Halterman, D. A., Jia, Y., & Martin, G. B. (1996). Initiation of plant disease resistance by physical interaction of AvrPto and Pto kinase. *Science*, *274*(5295), 2060–2063. https://doi.org/10.1126/science.274.5295.2060

Thines, M. (2014). Phylogeny and evolution of plant pathogenic oomycetes-a global overview. *European Journal of Plant Pathology*, *138*(3), 431–447. https://doi.org/10.1007/s10658-013-0366-5

Thines Marco, M., & Kamoun, S. (2010). Oomycete-plant coevolution: Recent advances and future prospects. *Current Opinion in Plant Biology*, *13*(4), 427–433. https://doi.org/10.1016/j.pbi.2010.04.001

Thirugnanasambantham, K., Durairaj, S., Saravanan, S., Karikalan, K., Muralidaran, S., & Islam, V. I. H. (2015). Role of Ethylene Response Transcription Factor (ERF) and Its Regulation in Response to Stress Encountered by Plants. *Plant Molecular Biology Reporter*, *33*(3), 347–357. https://doi.org/10.1007/s11105-014-0799-9

Van Damme, M., Huibers, R. P., Elberse, J., & Van Den Ackerveken, G. (2008). *Arabidopsis* DMR6 encodes a putative 2OG-Fe(II) oxygenase that is defense-associated but required for susceptibility to downy mildew. *Plant Journal*, *54*(5), 785–793. https://doi.org/10.1111/j.1365-313X.2008.03427.x

Van Den Ackerveken, G. F. J. M., Vossen, P., & De Wit, P. J. G. M. (1993). The AVR9 race-specific elicitor of Cladosporium fulvum is processed by endogenous and plant proteases. *Plant Physiology*, *103*(1), 91–96. https://doi.org/10.1104/pp.103.1.91

Van Der Biezen, E. A., & Jones, J. D. G. (1998). Plant disease-resistance proteins and the gene-for-gene concept. *Trends in Biochemical Sciences*, *23*(12), 454–456. https://doi.org/10.1016/S0968-0004(98)01311-5

Van Der Hoorn, R. A. L., & Kamoun, S. (2008). From guard to decoy: A new model for perception of plant pathogen effectors. *Plant Cell*, *20*(8), 2009–2017. https://doi.org/10.1105/tpc.108.060194

Varshney, A., Mohapatra, T., & Sharma, R. P. (2004). Development and validation of CAPS and AFLP markers for white rust resistance gene in *sis*. *Theoretical and Applied Genetics*, *109*(1), 153–159. https://doi.org/10.1007/s00122-004-1607-y

Vogel, J. P., Raab, T. K., Schiff, C., & Somerville, S. C. (2016). PMR6 , a Pectate Lyase-Like Gene Required for Powdery Mildew Susceptibility in *Arabidopsis* Author (s): John P . Vogel , Theodore K . Raab , Celine Schiff and Shauna C . Somerville Published by : American Society of Plant Biologists (ASPB) Stable URL : *Plant Cell*, *14*(9), 2095–2106. https://doi.org/10.1105/tpc.003509.sis

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0112963

Walsh, J. (2002). Pathogen profile. *Molecular Plant Pathology*, *3*, 289–300.

Wang, F., Wang, C., Liu, P., Lei, C., Hao, W., Gao, Y., Liu, Y. G., & Zhao, K. (2016). Enhanced rice blast resistance by CRISPR/ Cas9-Targeted mutagenesis of the ERF transcription factor gene OsERF922. *PLoS ONE*, *11*(4), 1–18. https://doi.org/10.1371/journal.pone.0154027

Wang, K. (2014). Agrobacterium protocols: Third edition. *Agrobacterium Protocols: Third Edition*, *1*, 1–368. https://doi.org/10.1007/978-1-4939-1695-5

Wang, X., Richards, J., Gross, T., Druka, A., Kleinhofs, A., Steffenson, B., Acevedo, M., & Brueggeman, R. (2013). The rpg4-mediated resistance to wheat stem rust (Puccinia graminis) in barley (Hordeum vulgare) requires Rpg5, a second NBS-LRR gene, and an actin depolymerization factor. *Molecular Plant-Microbe Interactions*, *26*(4), 407–418. https://doi.org/10.1094/MPMI-06-12-0146-R

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C., & Qiu, J. L. (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nature Biotechnology*, *32*(9), 947–951. https://doi.org/10.1038/nbt.2969

Warwick, S. I., & Al-Shehbaz, I. A. (2006). *Brassicaceae*: Chromosome number index and database on CD-Rom. *Plant Systematics and Evolution*, *259*(2–4), 237–248. https://doi.org/10.1007/s00606-006-0421-1

Wei, C., Chen, J., & Kuang, H. (2016). Dramatic number variation of r genes in solanaceae species accounted for by a few r gene subfamilies. *PLoS ONE*, *11*(2), 1–15. https://doi.org/10.1371/journal.pone.0148708

Weng, H., Pan, A., Yang, L., Zhang, C., Liu, Z., & Zhang, D. (2004). Estimating number of transgene copies in transgenic rapeseed by real-time PCR assay with HMG I/Y as an endogenous reference gene. *Plant Molecular Biology Reporter*, *22*(3), 289–300. https://doi.org/10.1007/BF02773139

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, *3*(10), 0–6. https://doi.org/10.1099/mgen.0.000132

Wick, R. R., Judd, L. M., & Holt, K. E. (2018). Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *BioRxiv*, 1–11. https://doi.org/10.1101/366526

Xie, K., & Yang, Y. (2013). RNA-Guided genome editing in plants using a CRISPR-Cas system. *Molecular Plant*, *6*(6), 1975–1983. https://doi.org/10.1093/mp/sst119

Yang, S., Zhang, X., Yue, J. X., Tian, D., & Chen, J. Q. (2008). Recent duplications dominate NBS-encoding gene expansion in two woody species. *Molecular Genetics and Genomics*, *280*(3), 187–198. https://doi.org/10.1007/s00438-008-0355-0

York, N. (2016). *Proceedings of the United Nations Security Council Meeting* (Issue March).

Yu, S., Zhang, F., Zhao, X., Yu, Y., & Zhang, D. (2011). Sequence-characterized amplified region and simple sequence repeat markers for identifying the major quantitative trait locus responsible for seedling resistance to downy mildew in Chinese cabbage (B*rassica rapa ssp. pekinensis*). *Plant Breeding*, *130*(5), 580–583. https://doi.org/10.1111/j.1439-0523.2011.01874.x

Zeyen, R. J., Carver, T. L. W., & Lyngkjær, M. F. (2002). Epidermal cell papillae. *The Powdery Mildews. A Comprehensive Treatise.*, *June*, 107–125.

Zhang, X., Wang, W., Guo, N., Zhang, Y., Bu, Y., Zhao, J., & Xing, H. (2018). Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC Genomics*, *19*(1), 1–12. https://doi.org/10.1186/s12864-018-4582-4

Zhang, Yi, Massel, K., Godwin, I. D., & Gao, C. (2019). Correction to: Applications and potential of genome editing in crop improvement. *Genome Biology*, *20*(1), 1–11. https://doi.org/10.1186/s13059-019-1622-6

Zhang, Yunwei, Bai, Y., Wu, G., Zou, S., Chen, Y., Gao, C., & Tang, D. (2017). Simultaneous modification of three homoeologs of TaEDR1 by genome editing enhances powdery mildew resistance in wheat. *Plant Journal*, *91*(4), 714–724. https://doi.org/10.1111/tpj.13599

Zipfel, C. (2008). Pattern-recognition receptors in plant innate immunity. *Current Opinion in Immunology*, *20*(1), 10–16. https://doi.org/10.1016/j.coi.2007.11.003

Appendix

Table 1. Flanking primers used to amplify candidate resistance genes within the BraA06 region mapped in *Brassica rapa* accessions EH_25 and EH_5, which confers resistance to race 2 isolates AcBj12 and AcBjDC.

Gene ID	Forward/Reverse primer	Sequence
BraA06g001890	F	CTTCAAAAGAAACCTAGCC
	R	TCATGTCCATCCCGACAAAGA
BraA06g002390	F	GTTGTACCTCATCTCATGTTA
	R	ACTGGAGAGAAGTGGAAAT
BraA06g003120	F	CGAAGACCGACCAATGACT
	R	GCCATTAACCAAGTTCAGA
BraA06g003420	F	TCTGTTACACCTTGGAGGA
	R	GAGAGGATACACTCACACA
BraA06g003430	F	GGCGGTTTAACCCTAGTAT
	R	CAACTGGGGATTCAGAGCC
BraA06g003690	F	ACTCCTGTTCGTGTCTCTGAA
	R	GTTCATTGGCTCTATGCACGG
BraA06g003770	F	GCATATATTTGTATCCGTGCA
	R	ACAAACTGGAGAGAAGTGG

Script 1. Produce physical maps shown in Chapter 2.

```
library(tidyverse)

### Import data, with the appropriate columns labelled as below.
setwd("/Users/u1590355/Desktop/Master_Folder/Data_Sets/GBS_Raw_data/BJ12/R sorted output for
 Excel input")
Data <- read_csv("BJ12_Dom_C.csv")

Chr <- Data$Chr
Pos <- Data$Position
type <- Data$Isolate
Pop <- Data$Population
Chr_sizes <-
 c(31028331,32791447,25129783,21622048,28517042,29418425,27501058,23223082,44954983,19932
 799)

data<-data.frame(chromosome=paste0("chr", c(1,2,3,4,5,6,7,8,9,10.0)),size= Chr_sizes ,stringsAsFactors =
 FALSE)
data$chromosome<-factor(data$chromosome, levels = data$chromosome)
SNP<-data.frame(chromosome= Chr,Position= Pos,Type= type,labels= Pop)


### Plot chromosome map with samples labelled
p <- ggplot(data=data, aes(x=chromosome, y=size)) +
geom_segment(data = data, aes(x = chromosome, xend = chromosome,  y = 0, yend = size), lineend =
 "round", color = "lightgrey", size = 5)  +
theme_minimal()
```