Next generation genomics for improved disease resistance in bread wheat

Clémence Marchal

A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

John Innes Centre

Norwich

November 2019

Abstract

Crop diseases reduce wheat yields by ~25% globally and thus pose a major threat to global food security. Yellow (stripe) rust caused by *Puccinia striiformis* f. sp. *tritici*, is distributed worldwide and is currently the most globally damaging cereal rust. Despite over 80 designated yellow rust resistance genes (*Yr*) in wheat, few have been cloned.

Using mutational resistance gene enrichment sequencing (MutRenSeq) we successfully cloned three non-canonical BED domain containing Nucleotide-binding and Leucine-rich Repeat proteins (BED-NLRs) in wheat that confer different resistance spectra to yellow rust: *Yr7*, *Yr5* and *YrSP*. We showed that all three genes are genetically linked and *Yr5* is distinct from *Yr7*, whereas *YrSP* is a truncated allele of *Yr5* with 99.8% sequence identity. We demonstrated that a single amino-acid change in the BED domain of *Yr7* was sufficient to lead to a loss of resistance. Additionally, Yr5 and YrSP BED domains are identical and there is only one amino-acid polymorphism between Yr7 and Yr5/YrSP BED domains. We thus hypothesized that recognition specificity is not solely governed by the BED domain.

Given the presence of integrated BED domains, we asked whether their mode of action would be similar to what was proposed in the 'integrated decoy' model. To test this hypothesis, we combined comparative genomics and neighbour-net analyses to determine whether BED domain from BED-NLRs are sequence-related to certain BED-domains from other BED-containing proteins. Additionally, we set-up transient expression assays in *Nicotiana benthamiana* to investigate the ability of *Yr7* and *Yr7* variants to trigger cell-death in this heterologous system. Together these results provide novel insights into the mode of action of BED-NLRs in disease resistance in crops.

Acknowledgments

I could not have completed my PhD without the support and guidance of many people and I would like to express my sincere gratitude with the following words.

I first would like to thank my supervisor, Cristobal Uauy for his continuous support and guidance during these four years. Thank you for introducing me to wheat and the wheat community. I am very grateful for your advice, from troubleshooting experiments to planning the next steps of my scientific career. I would like to thank my co-supervisor Brande Wulff for his insight and encouragements throughout my PhD project and for giving me the opportunity to learn bioinformatics within his group. I would also like to thank my co-supervisor Simon Berry for his support and guidance during my PhD thesis and for introducing me to the breeding side of the project and showing me its direct applications.

I would like to thank Group Limagrain UK for funding my PhD project and the John Innes Centre for hosting me. I could not have asked for a better environment to pursue this project.

Many people contributed to this project and I would like to thank Jianping Zhang, Evans Lagudah, Peng Zhang, Robert MacIntosh and Lesley Boyd for joining forces and collaborating on the *Yr* gene cloning project. A special thanks to Sreya, Burkhard and Ricardo who have been of a tremendous help during my bioinformatics training. I also would like to thank Paul Fenwick for helping with the pathology assays and sharing with me his great knowledge of wheat diseases. I am very grateful to Sophien Kamoun for his insight and the time I spent in his lab optimising the molecular biology part of my project. I would like to thank Mark Smedley to talk me through Golden Gate cloning and Sadiye Hayta for the transgenic wheat lines. I am also very grateful to Damian, Sophie, Lewis, Oscar and Tim from horticultural services for their help managing and tending to all these wheat plants, Richard from the genotyping platform for his help with KASP assays and Mark Youles from Synbio for his help with the Golden Gate modules.

I have been extremely lucky during my PhD to be part of two brilliant labs! Thank you so much all members of the Uauy and Wulff lab for your never-ending support and help. Thank you to my officemates Sreya, Amber, Sanu, Tom, Ngoni Guotai and Kumar for your help, laughter and continuous supply of cakes and biscuits! A special thanks to Sreya, I could always count on you for moral support in and outside the lab. I would also like to thank Philippa, Olu, Nikolai, Sophie and Jemima for their help settling in the lab and their invaluable tips! You are an amazing group of people to work with and you create a fun and collaborative atmosphere in the lab and the office.

I am grateful to have met many great friends in Norwich, with whom I shared many fun times, including Antoine, Pierre, Myriam, Émilie, Adeline and my former housemates Anna, Juan-Carlos, Neftaly, Ruth, Roland, Zigmunds and Alberto, thank you all! Special

thanks to my running buddies Bihai and Ricardo, it has been great training with you both! Thank you Helen for sharing with me horse riding lessons at Nine Acres.

I would also like thank my friends and colleagues who pre-date the PhD. A special thanks to my MSc internship supervisor Seb for believing in me and inspiring me to embark on a PhD! I am very grateful to Ben and Aurore for hosting me when I first arrived in Norwich and helping me settling in. Thank you Cécile, with whom I share many fond memories from my MSc internship, for keeping in touch and giving me PhD tips, Team PFF! Thank you as well to my dearest friend Claire for her never-ending moral support and for sticking around all these years since high-school.

A huge thanks to Thorsten, for your love and both scientific insight and moral support. Thank you also for your patience during the writing ☺.

Lastly, I would like to thank my family for their never-ending love and support. Thank you for continuously encouraging me despite my terrible communication, especially in the last weeks. Thank you, Grandpa and Grandma, for sending me food survival kits (saucisson and cheese are always welcomed!) and being only one Skype call away. Thank you so much Mum for your daily support messages and photos of the dogs, gecko, and, sometimes, Titi and Dad. They kept me going despite being away from you all. Thank you, Mum and Dad, for believing in me since day one.

Table of figures

Table of tables

Table of Appendices

List of abbreviations

APR: Adult Plant Resistance

BED domain: named after **BE**AF (boundary element-associated factor) and **D**REF (transcription factor) from *Drosophila*. zinc-finger DNA-binding domain

BLAST: Basic Local Alignment Search Tool

BSA: Bulk Segregant Analysis

BUSCO: Benchmarking Universal Single-Copy Orthologs

CC domain: Coiled-coil domain

Co-IP: Co-immunoprecipitation

EMS: Ethyl methanesulfonate

ETI: Effector-Triggered Immunity

ETS: Effector-Triggered Susceptibility

HA: Human influenza hemagglutinin

HMA: heavy-metal-associated domain

HMMER: software for sequence analysis based on Hidden Markov Models

HR: Hypersensitive response

KASP: Kompetitive allele specific PCR

Lr: Leaf rust resistance

MAS: Marker-Assisted Selection

Mla: Mildew Locus A (*Hordeum vulgare*)

MUMmer: Software for sequence alignment based on Maximal Unique Matches

MutRenSeq: Mutational genomics coupled with Resistance gene enrichment Sequencing

NIL: Near Isogenic Line

Continued next page

List of abbreviations continued

NLR protein: Nucleotide-Binding Leucine Rich Repeat protein (or NB-LRR)

NLS: Nuclear Localisation Signal

PAMPs: pathogen-associated molecular patterns

PRRs: pathogen recognition receptors

Pik-1/2: *Pyricularia oryzae* resistance k 1/2 (*Oryza sativa*)

Pst: *Puccinia striiformis* f.sp *tritici*

RATX1: Related to ATX1, a copper transport protein of *Saccharomyces cerevisiae*

RLK: Receptor-like Kinase

RLP: Receptor-like Protein

RGA 4/5: *R*-gene analogue 4/5 (*Oryza sativa*)

RPM1: Resistance to Pseudomonas syringae protein 3 (RPS3) (Arabidopsis)

RPS4: Resistance to *Pseudomonas syringae* 4 (Arabidopsis)

RRS1: Resistance to *Ralstonia solanacearun* 1 (Arabidopsis)

SNP: Single nucleotide polymorphism

Sr: Stem rust resistance

TILLING: Targeting Induced Local Lesions in Genomes

TIR domain: Toll/Interleukine-1 receptor domain

WRKY domain: DNA-binding domain found in the WRKY transcription factor family (characterised by a WRKYGQK motif)

WT: Wild-type

Yr: Yellow rust resistance

YFP: Yellow fluorescent protein

1. General Introduction

1.1. Wheat is a crop of global importance

1.1.1. The challenge of meeting food demand with an increasing population

Pardey et al., 2014[1] showed that the global food demand per person per day increased from an average of 2250 kilocalories (kcal) in the early 1960s, to ~2880 kcal in 2015. A more recent meta-analysis regrouping 22 independent studies showed that the average predicted food demand per capita in 2050 will be 3250 kcal/day (95% CI = 3176–3324) and would rise to 3527 kcal/day (95% CI = 3290–3763) in 2100[2]. Although the studies were variable in terms of model methods and data sources, the trend points towards an increase of the global food demand from present day to 2100[2].

Although the global population growth rate is declining since the 1970s, the total population is forecast to hit 9.8 billion by 2050[3]. This paired with the increased global food demand per person discussed above raises concerns about how food demands will be met and what will be the environmental impact[4]. There is no straight forward solution to this and a combination of different options will be required, from sustainably increasing productivity to adapting diets (discussed in Godfray et al., 2010[5]). It is also important to note that access to food remains a major issue in some regions of the world and that the prevalence of undernourished people is rising again since 2015 and was estimated at 10.8 % in 2018[6].

Increasing productivity in a sustainable manner is one of the measures needed to tackle the increasing food demand by the global population.

1.1.2. Global wheat production

Wheat and its derived products represent on average 19.8 % of the world's total calorie intake and 22.3 % of the world's total protein intake (calculated from FAO data acquired at http://www.fao.org/faostat/en/#data/CC, Figure 1-1). Despite continuously increasing since the 1960s with the Green Revolution, cereal yields are now plateauing and predicted to be insufficient to meet food demand in 2050[7,8]. Does it mean that we have reached the maximum yield potential for all crops? Other reasons may explain this, including climate change, land degradation over the years or location of production areas in poor soils and climate conditions[7]. Interestingly a study of wheat yield in France showed that the potential yield gain achieved through genetic progress was counteracted by climate becoming more unfavourable to cereal yields[9]. The authors also pointed out that they could not rule out agronomic causes related to policy and economy as contributing factors. There is thus still room for improvement regarding wheat yields.

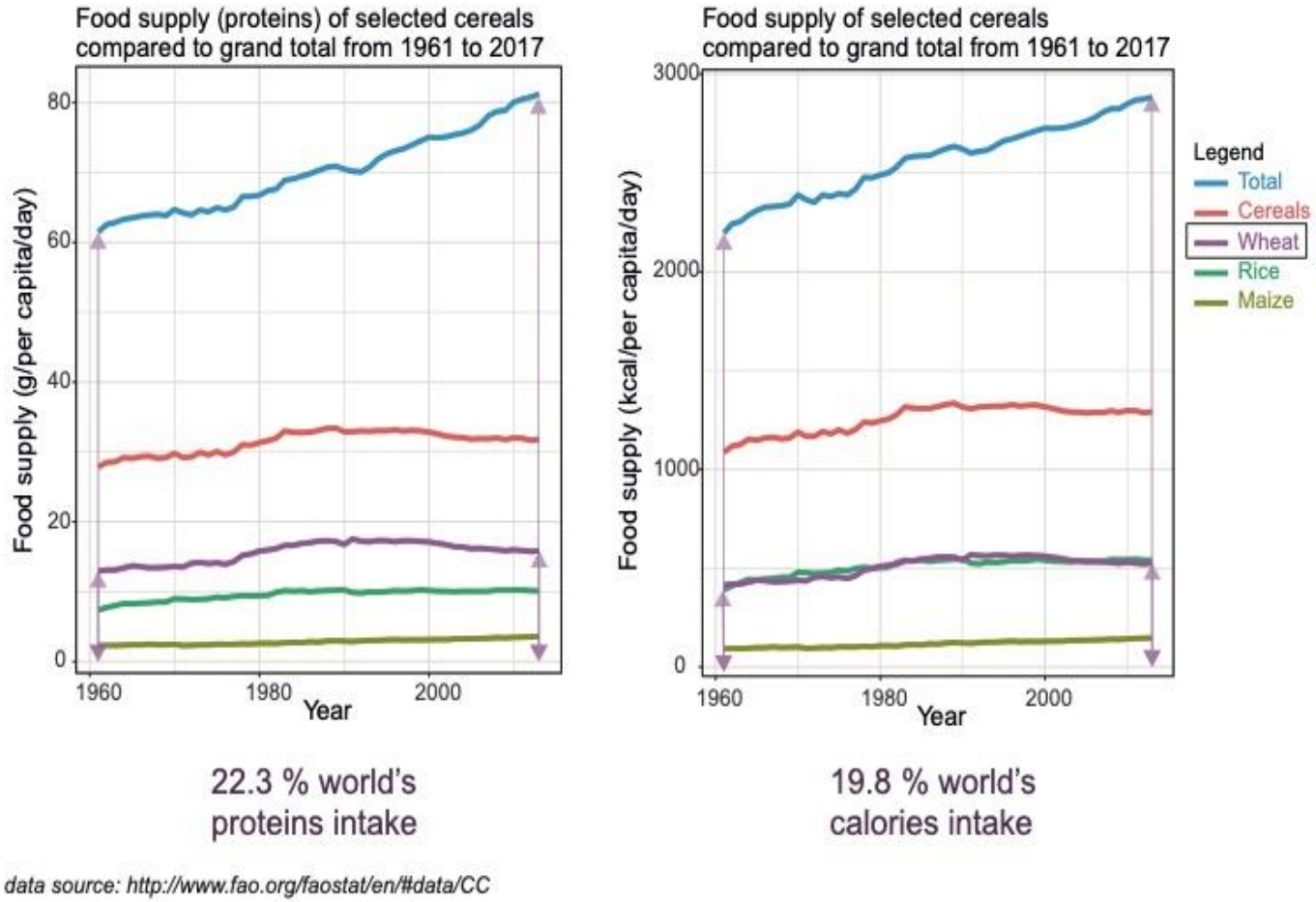


Figure 1-1. Proportion of selected cereals including wheat in total food supply from 1961 to 2017.
Left: Proportion of selected cereals in total proteins/person/day from 1961 to 2017. Right: Proportion of selected cereals in total kcal/person/day from 1961 to 2017. Data acquired from http://www.fao.org/faostat/en/#data/CC.

1.2. The features of wheat

1.2.1. The origin of cultivated wheat

Development of agriculture was a major step in human history. It marked a profound change of lifestyle from hunter gatherer to sedentary farmers, known as the Neolithic Revolution[10]. Cereals began to be cultivated for their seeds and subsequently the domestication process started to make crops easier to harvest. Einkorn wheat (*Triticum monococcum*, genomes A^mA^m) was among the first domesticated cereals[11] and its domestication occurred in southeast Turkey[12].

Modern cultivated wheat is allopolyploid, that is a polyploid that has arisen through the hybridisation of chromosomes from different species. Common wheat is the result of two natural genome hybridisations with closely related *Aegilops* species (Figure 1-2)[13]. The first hybridization event occurred approximately 400,000 years ago between two diploid grass species (*Triticum urartu* (A^uA^u) and an unknown member of the Sitopsis family (BB) that includes *Ae. speltoides*, *Ae. longissima*, *Ae. sharonensis*, *Ae. searsii*, and *Ae. bicornis*[14]. The B genome donor is hypothesized to be closely related to *Ae. speltoides*[14] (genome SS). This gave rise to tetraploid wild emmer *Triticum turgidum* ssp. *dicoccoides* (AABB). The domestication of wild emmer gave rise to emmer wheat (*Triticum turgidum* ssp. *dicoccon*, AABB) and was the first step that ultimately resulted in the evolution of free-threshing tetraploid durum wheat (*T. turgidum* ssp. *durum*, AABB)[15]. Our modern pasta wheat *T. durum* (AABB) originated from *T. turgidum* ssp. *durum* (Figure 1-2).

The second hybridization step occurred 10,000 years ago between emmer wheat and *Aegilops tauschii* (DD)[13]. The spread of emmer wheat cultivation in the growth area of

Ae. tauschii facilitated this hybridization[13]. This formed common hexaploid wheat *Triticum aestivum* (AABBDD) that gave rise to our modern bread wheat varieties *Triticum aestivum* ssp. *aestivum*[16] (Figure 1-2). Cultivation of wheat spread worldwide due to its ability to grow at a wide range of climatic conditions and high yield.

There is evidence for other hybridization events having occurred between hexaploid wheat and emmer. For example, *Triticum aestivum* ssp. *spelta* (spelt) may have risen from the hybridization between free-threshing hexaploid wheat and emmer[17] (Figure 1-2). Because European and Asian spelt are distant, it is likely that their ancestral hexaploid wheat may have encountered different tetraploid wheats depending on the growth location. This generated different introgression and hybridization events, leading to distant spelt wheat in Europe and Asia[18]. Spelt wheat is the source of *Yr5*, one of the disease resistance genes studied in this thesis.

The focus of this work is bread wheat (*Triticum aestivum* ssp. *aestivum*, AABBDD). Bread wheat accounts for 95 % of total cultivated wheat. We will thus refer to it as 'wheat' in this thesis, unless stated otherwise.

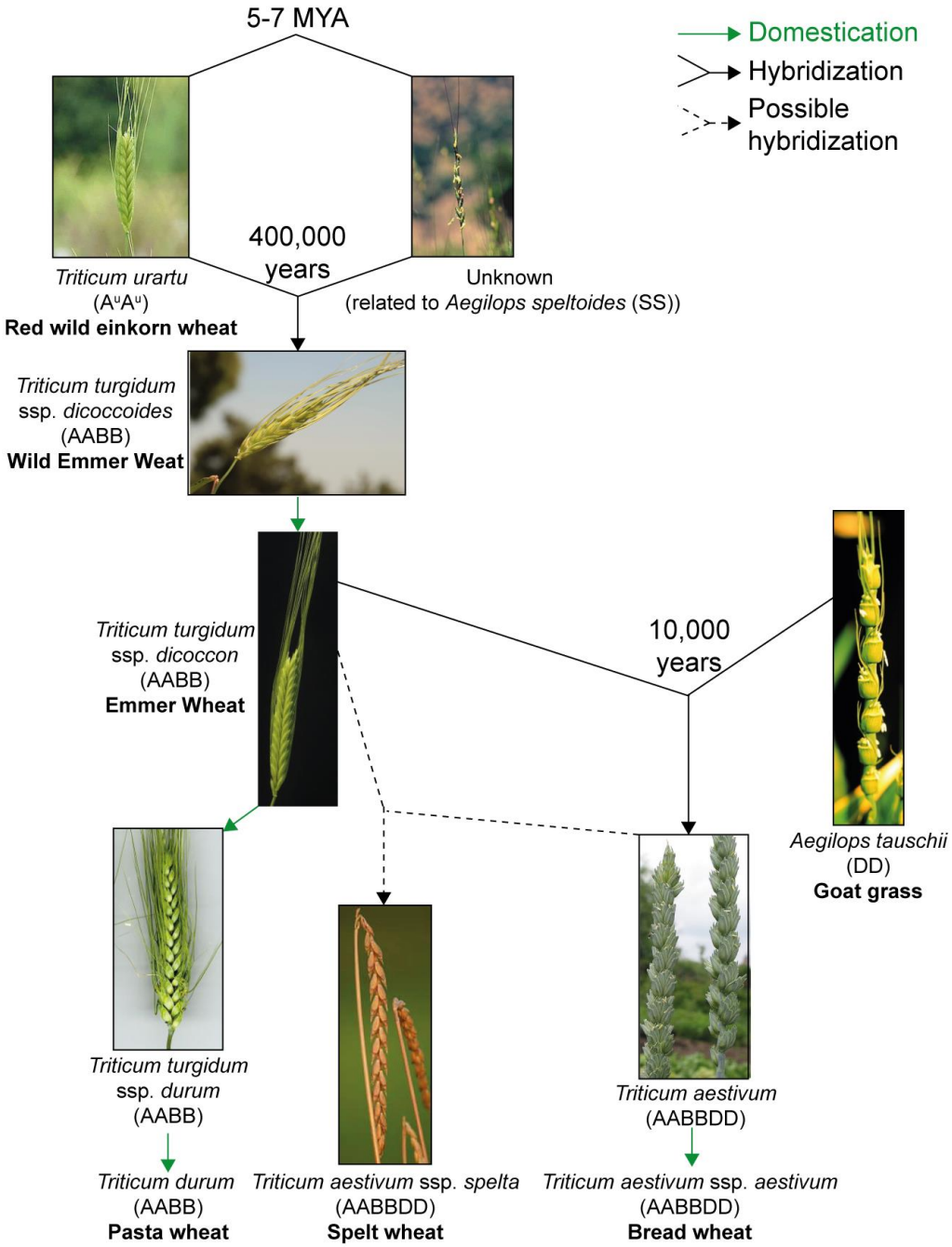


Figure 1-2. Diagram representing the evolution of bread, pasta and spelt wheat as described in section 1.2.1

1.2.2. Wheat domestication

Domestication refers to the process of artificially selecting plants to increase their suitability to human requirements including taste, yield, storage, and cultivation practices. For example, we discussed in the section above that the domestication of wild emmer resulted in the evolution of free-threshing tetraploid durum wheat[15]. This made

separation of the grain from the spikelet less labour intensive. Additionally, reduction of spike-shattering was another very important trait in wheat domestication because it enabled easy harvesting of grain via preventing them to shatter on the ground. This trait is encoded by the brittle rachis (*Br*) genes, located on the short arms of chromosome 3A and 3B[19]. Seed size, reduced tiller number and a more erect growth habit were also among the numerous traits selected thorough domestication[20]. We stated in section 1.1.1 that the Green Revolution allowed for increased harvest index, the ratio of grain yield to the above ground tissue at maturity[21]. The semi-dwarf *Rht* alleles introduced from Japanese cultivars led to reduced stem length and were among the selected traits that allowed for increased yields in wheat over this period[21].

Domestication led to a reduction in genetic diversity in modern varieties relative to wild progenitor species. This is attributed to an initial dramatic reduction in population size termed the "domestication bottleneck," followed by an expansion in population size[22]. For example, domestication of tetraploid wheat was accompanied by a loss of genetic diversity in domesticated varieties and a shift in allele frequencies toward more common alleles[23]. Although homogeneity is advantageous for cultivation, such changes are disadvantageous for plants in the wild environment. Some strategies used in breeding involve going back to the wild relatives of wheat to identify favourable traits such as disease resistance.

1.3. Yellow rust disease of wheat

1.3.1. Cereal rust diseases are among the major biotic constraints applied to wheat yield

Although wheat is a successful crop that is grown worldwide, maintaining a sustainable wheat yield remains a challenge in certain environments. This is mostly due to the numerous pathogens/pests which are responsible for about 50 % of the global wheat yield losses[24]. Cereal rusts have historically been among the major biotic constraints in world wheat production[25]. A recent study on the effect of pathogens on the global yield of different crops showed that cereal rusts belong to the top 10 diseases causing most of the yield loss in wheat, a list which also included Fusarium Head Blight (FHB), Tritici and Spot Blotch, and powdery mildew[26]. Additionally, rust epidemics can be devastating locally and represent a heavy burden on local farmers who can lose up to 100 % of their harvest due to these diseases.

One of the most dramatic examples is the Ug99 epidemics that occurred in Africa and Middle-East in 1999 (first detected in Uganda in 1998). This stem rust pathogen race threatened wheat production worldwide because 90 % of the varieties were found to be susceptible at that time[27]. In 2005, Nobel Laureate Dr. Norman E. Borlaug raised the alarm about the serious threat Ug99 could pose to food security if proper actions were not taken. The wheat community and donor organisations responded positively and coordinated research and development projects to respond to the Ug99 epidemics. This led to the creation of the Borlaug Global Rust Initiative (BGRI, http://www.globalrust.org). BGRI is still active and is now a wide network bringing together scientists, industries and funding bodies focusing research effort on developing sustainable resistance against cereal rusts.

1.3.2. The causal agents of rust diseases in wheat

1.3.2.1. The three main rust diseases occurring in wheat

Wheat rust pathogens belong to the genus *Puccinia,* family *Pucciniaceae*, order Uredinales and class Basidiomycetes. There are three main wheat rust diseases: wheat stem rust is caused by *P. graminis* f. sp *tritici* (*Pgt*), wheat leaf rust by *P. triticina* (*Pt*) and wheat stripe rust by *P. striiformis* f. sp *tritici* (Pst). They are also commonly named black, brown and yellow rust, respectively, given the induced symptoms on the wheat plant (Figure 1-3). The yield loss is mainly due to the production of pustules, which reduces the photosynthetic capacity of the host plant.

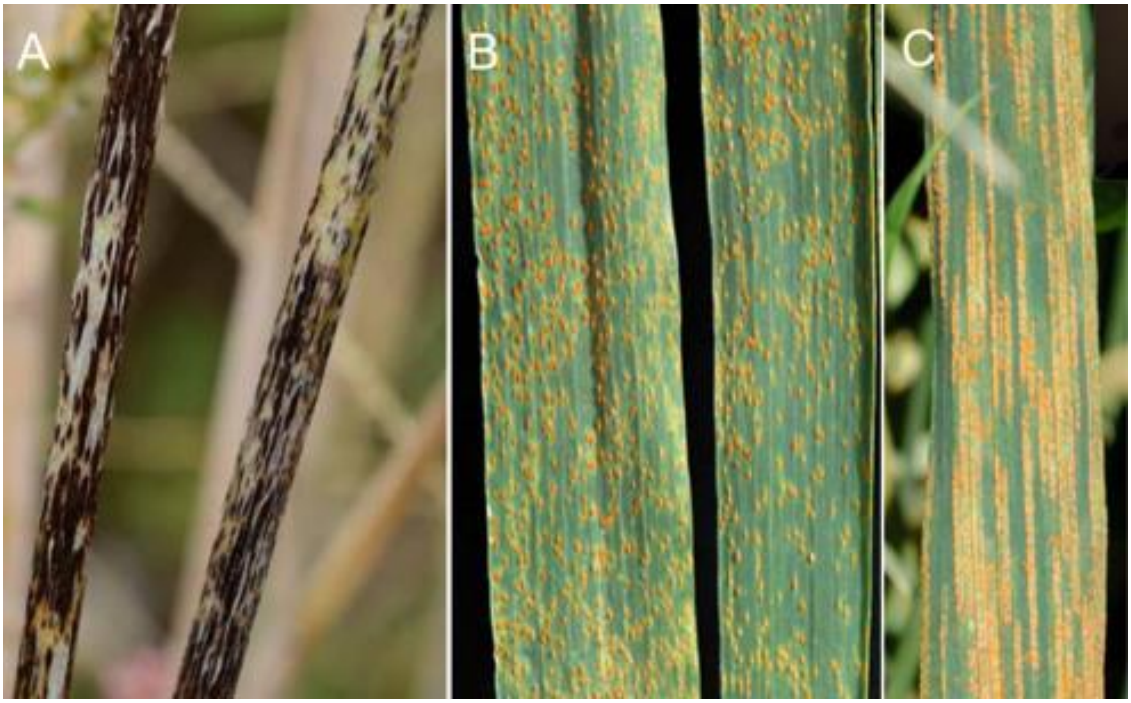


Figure 1-3. Pictures showing the wheat leaf symptoms corresponding to the three main rust diseases.
(A) Stem rust. (B) Leaf rust. (C) Stripe rust. Pictures from https://www.ars.usda.gov/midwest-area/stpaul/cereal-disease-lab/docs/cereal-rusts/

Rusts pathogens are specific obligate parasites that interact with wheat, among other ways, in a gene-for-gene relationship[28] (further discussed in section 1.3.3). Because of

this specificity, the virulence of rust fungi against cereal resistance is highly diverse and results in the existence of many different pathogenic races. Cereal rusts can be disseminated thousands of kilometres across continents and oceans by wind in the form of clonally produced dikaryotic urediniospores[29] (see section below). Foreign races can therefore be introduced in areas far removed from the sites of their original detection and thus can have terrible consequences on the locally cultivated crops, which are not adapted to these incoming races. Until recently, stem rust (caused by Ug99 *Pgt* race among other isolates) was considered more damaging regarding wheat yield loss than stripe and leaf rust. However, given its geographical extent and the associated production losses, stripe rust has been suggested as the most damaging of all the cereal rusts nowadays[30,31]. At least 4.70 million tons are annually lost due to this pathogen (being equivalent to a US$840 million annual loss) and 88% of the world's wheat crop production is seasonally vulnerable to stripe rust[32].

1.3.2.2. *Pst has a complex life cycle that allows for rapid adaptation to the host*

Over the last century, it was assumed that *Pst* was is a macrocyclic, heteroecious fungus based on similarities with other cereal rust fungi. The life cycle and biology of this fungus have been reviewed elsewhere[33–35] and we will provide a summary in this section and in Figure 1-4.

Pst is heteroecious because its life cycle is completed on two phylogenetically different hosts, wheat and different subspecies of *Berberis* (*B. chinensis*, *B. holstii*, *B. koreana*, *B. vulgaris*). Wheat is the primary host where *Pst* asexually multiplies, whereas the sexual recombination occurs on the alternate host *Berberis*. *Berberis* was only identified as *Pst*'s alternate host a decade ago[36]. The macrocyclic character of *Pst* life cycle is defined by the five different forms that the fungus undergoes to complete it:

- Urediniospores are produced by the Uredinia on wheat leaves. These spores are dikaryotic (n +n) and constitutes the form that is responsible for rust epidemics.
- At the end of the wheat growing season, telia form on the leaf epidermis and these structures produce teliospores (2n).
- Teliospores germinate and undergo karyogamy and meiosis to produce basidiospores (n) that infect the alternate host *Berberis.*
- Basidiospores form Pycnia on the upper side of *Berberis* leaves, that lead to disease symptoms on the alternate host and produce pycniospores (n).
- Pycniospores form Aecia clusters on the lower side of *Berberis* leaves and produces aeciospores (n + n) that will infect wheat.
- Aeciospores form Uredinia on wheat leaves and the cycle is complete.

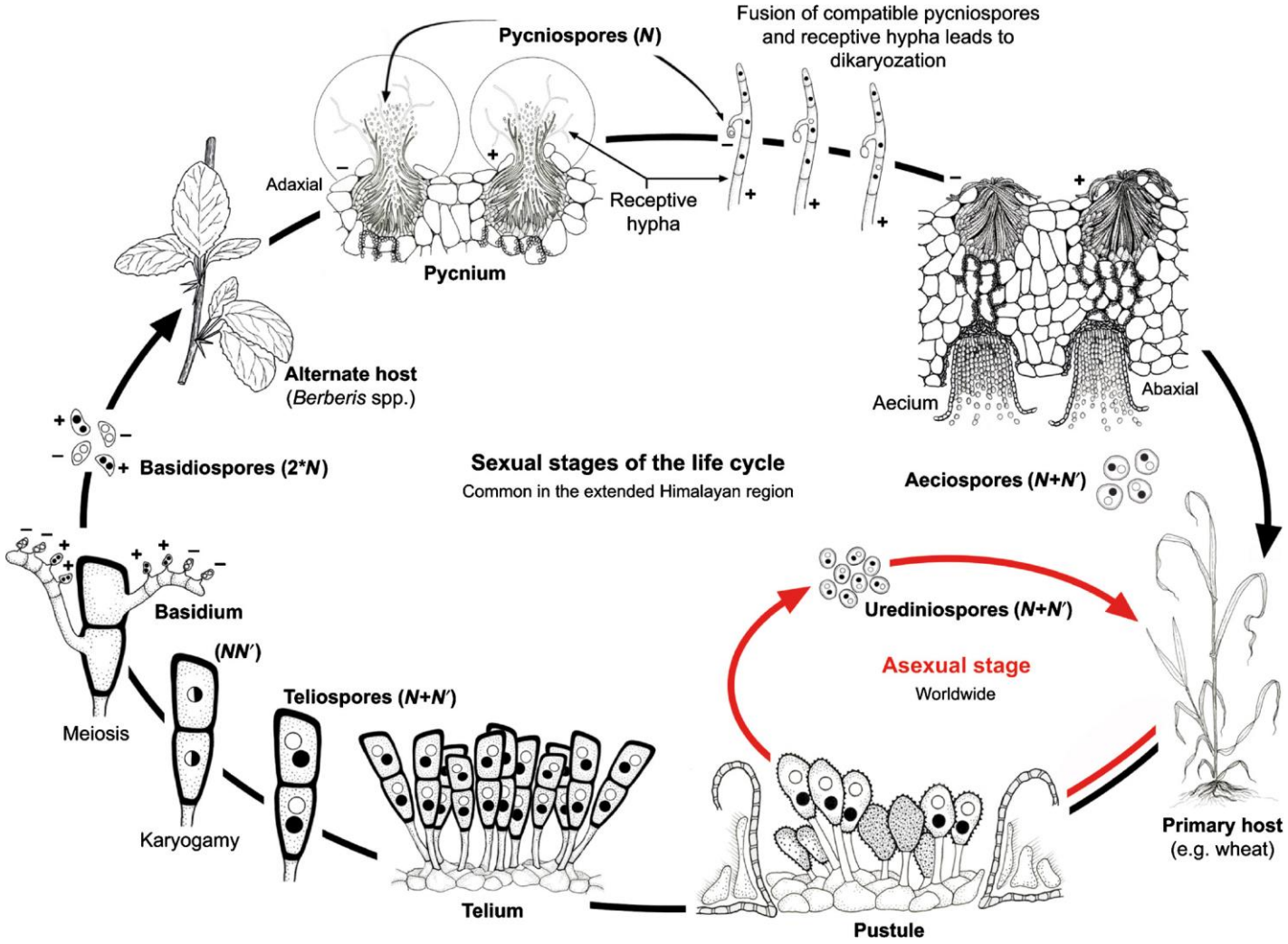


Figure 1-4. Life cycle of *Puccinia striformis* f. sp. *tritici*.
Figure from Schwessinger, 2017[37].

The complex life cycle of *Pst* enables genetic diversity and rapid adaptation[34]. Although sexual recombination only occurs on *Berberis* and is a major source of genetic diversity, it has been shown that each nucleus in the urediniospore phase (dikaryotic) accumulates mutations independently[35]. This leads to high heterozygosity rates between the nuclei. Additionally, it has been suggested that somatic hybridization, that is asexual exchange of genetic material, could generate genetic diversity during the asexual stage[38]. Very recently, Li et al., 2019[39] demonstrated that the Ug99 *Pgt* lineage arose by somatic hybridisation and nuclear exchange between dikaryons. Similar mechanism could thus occur in *Pst*. This combined with the consequent number of spores produced at each stage and the ability of *Pst* spores to travel across long distances leads to millions of potential variations at each genetic locus within one growing season (reviewed in Schwessinger, 2016[34]). Given that we grow wheat as monoculture, it is not surprising to observe new virulent variants able to infect wheat fields globally in a short period of time[40]. This consists a major challenge when it comes to developing sustainable resistance.

There are three main ways to control rust diseases in cereals:

- Agronomy through reducing the pathogen alternative host population that is necessary for the pathogen sexual recombination, thus decreasing the risk of a new virulent race to emerge. This has been successfully applied on *Berberis vulgaris* in Europe and North America to control stem rust [41]. Such management is however difficult to set up. Another alternative would also be to avoid continuous wheat cultivation across the year in certain areas. Indeed, the pathogen can only remain on living plants. Thus, if the host is cultivated all year round in a specific location, the pathogen can rapidly infect the new seedlings.

- Chemicals are generally efficient and widely used to control rust diseases, but they are not environmentally friendly and the pathogen can develop resistance to fungicides. Moreover, farmers from developing countries can rarely afford to use them because of their high cost and suffer from lack of timely access.

- Genetic resistance is less harmful for the environment and growing a resistant wheat variety is less expensive than having to spray susceptible fields several times per season. It has however to be used in a well-reasoned way to avoid it being defeated by the pathogen given its ability to rapidly adapt.

Within the frame of the presented work, we focussed on genetic resistance against *Pst* and will thus provide more information about types and sources of resistances that can be used in wheat.

1.3.3. Two types of resistance and sources of resistance

Genetic resistance comprises a range of plant phenotypic responses to avoid or reduce pathogen colonisation. These responses can occur at different growth stages of the plant and can be more or less specific to different variants of the pathogen. To simplify this, the diverse set of responses has been traditionally classified into two broad categories: seedling and adult plant resistances. The name 'seedling' refers to the growth stage when this type of resistance is assayed despite the resistance being usually observed across the whole plant life cycle. On the other hand, adult plant resistance (APR) is not present at these early stages but manifests itself later on during plant development. It has been assumed that seedling resistance is specific to a certain pathogen race and adult plant resistance would have a broader spectrum resistance. Finally, it has commonly been accepted that seedling resistance is linked to a specific gene family named *R*-genes (for

Resistance-genes) due to its characteristics, whereas the broad-spectrum resistance inherent to adult plant resistance less likely to rely on a single gene family. We will describe both kinds of resistance in the following section and point out that this separation is not always obvious.

1.3.3.1. *Seedling resistance*

R-genes involved in seedling resistance mostly conform to Flor's gene for gene hypothesis[28], suggesting that two key genes are involved to allow resistance expression: the *R*-gene in the host recognizing the corresponding avirulence effector gene (*Avr*) in the rust pathogen. This also explains the strain specificity characterising *R*-genes. Most of the *R*-genes belong to the NLR family (or NBS-LRRs, Nucleotide-Binding Leucine Rich Repeat protein), which displays a characteristic domain pattern: a Toll/Interleukine-1 receptor (TIR) or a Coiled-coiled (CC) domain on the 5' end, followed by a Nucleotide Binding Site (NBS) domain and finally a succession of Leucine-Rich Repeats (LRR) on the 3' end[42] (Figure 1-5[43]). It had been suggested that the TIR and CC domains could be involved in interactions between two NLRs working in pair and in the response signalling[44,45], whereas the NBS domain allow ATP and/or GTP hydrolysis. The various roles of the LRR motifs, including effector recognition was reviewed elsewhere[46]. The significance of this specific domain pattern for studying this family will be discussed in section 1.6.5.

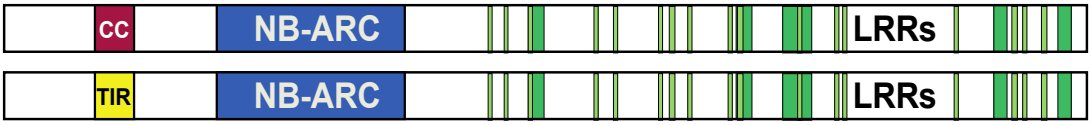


Figure 1-5. Illustration of the two main classes of NLR.
Coiled-Coil NLR (CNL, top) and TIR-NLR (TNL, bottom). Coiled-coil domain is shown in red, TIR domain in yellow, NB-ARC in blue and LRR regions in green.

The NLRs set, namely NLRome, of various plant have been studied since this gene family has been linked to disease resistance, including Arabidopsis, rice, poplar, potato, tomato or more recently cassava[47–52]. NLR loci are often organized into clusters of diverse sizes (first reviewed in Michelmore and Meyers, 1998[53]). For example, a Quantitative Trait Loci (QTL, described in section 1.6.1) linked to disease resistance located on the long arm of chromosome 2B of wheat possess at least 6 resistance genes to yellow rust (namely *Yr5*, *Yr7*, *Yr43*, *Yr44*, *Yr53* and *Yr 'Spaldings Prolific'*)[54]. Evidence was additionally found regarding *Yr5* and *Yr7* being allelic variants, or very closely linked[55]. Studying the gene organization and the allelic variation across the wheat NLRome would provide insights regarding their relation and evolution.

As discussed above, resistance conferred by *R*-genes is "all-stages" and displays a strong resistance phenotype characterized by a locally induced cell death, namely hypersensitive response (HR). This allows their rapid detection in glasshouses tests and thus makes selection simple and economical, which is advantageous in breeding programmes. However, one single mutation in the pathogen *Avr* gene can lead to the loss of recognition and therefore the loss of resistance in the host. This thus applies a high selection pressure on the virulent pathogen strains, which often overcome the *R*-gene within a few years after its first release (discussed above in section 1.3.2.2). We provide two example of defeated yellow rust resistance genes in wheat in Chapter 2. Although most of the *R*-genes have been defeated when deployed alone, they have been used with considerable success to control rust in many parts of the world by deploying varieties carrying several *R*-genes effective against most of the local rust races[56].

1.3.3.2. *Adult Plant Resistance*

Unlike *R*-genes, APR genes express rust resistance phenotypes at adult stages only. They are characterized by a partial resistance, with lesser and slower pathogen growth without

any noticeable HR. Consequently, APR is mostly detected and selected in the field and is assumed to apply a less stringent selection pressure on virulent isolates. However, as *R*-genes phenotypes are very strong and can be detected at all stages of the plant growth, they could potentially mask effective APR genes. This hinders the ability to identify and fine map APR genes thereby limiting their targeted use in wheat breeding.

Combining APR genes can lead to 'near immunity' in adult field grown plants[57]. Moreover some APR genes are able to enhance the level of *R*-gene resistance, such as *Sr2* which is the best-known APR gene in wheat and was genetically defined and mapped to chromosome arm 3BS[58]. *Lr34* is another well-studied APR gene in wheat and is also able to improve the effectiveness of *R*-genes[59,60]. These studies show that combining *R* and APR genes is a very promising strategy for rust disease control in wheat.

Isolation of APR genes is difficult given the partial resistance response. However, the phenotypic effects are often strong enough to allow genetic fine mapping and identification of loss of function mutants, which is important for map-based cloning. Two APR genes have been successfully cloned: *Lr34* (*Yr28, Pm8, Sr57*) encodes an ABC transporter, whose abscisic acid is a substrate that potentially have a role in transcriptional response of *Lr34*-resistant plants[61,62]; *Yr36* confers broad spectrum resistance to stripe rust and encodes a protein kinase with a lipid-binding domain, it has thus been renamed WKS1 for Wheat Kinase START1[63]. None of these APR genes belong to the NLR gene family. Their roles had not been clearly defined yet, but studies gave insights regarding the mechanisms they might be involved in. For example WKS1 has been shown to phosphorylate a thylakoid-associated ascorbate peroxidase and reduce its ability to detoxify peroxides in the chloroplast, potentially promoting cell-death and thus limiting pathogen proliferation in plant tissues[64]. More recent work also showed that

WKS1 interacts with and phosphorylates an extrinsic member of photosystem II named PsbO to reduce photosynthesis and regulate leaf chlorosis in conferring *Pst* resistance[65]. Studying further APR genes and uncovering the mechanism of their resistance would thus provide insights about how it could interact with seedling resistance, allowing working on an efficient combination in commercial varieties.

1.3.3.3. The distinction between R and APR genes appears to be less clear than previously thought

R-genes are widely regarded as belonging to the NLRs gene family, being race-specific and conferring resistance at early stages whereas APR genes are described as conferring a broad-spectrum resistance only at adult stage and not belonging to a specific gene family. However, several examples suggest that the boundaries between *R* and APR genes are more diffuse. Thus, this arbitrary classification might over-simplify the complexity of genetic disease resistance. Additionally, *R*-gene terminology sometimes include receptors of PAMPs (Pathogen-Associated Molecular Patterns, section 1.4), as some examples showed that they also could confer full resistance[66].

Evidence of race specific APR genes has been reported for stripe rust. Four QTL from a recombinant inbred line population (‘Camp Rémy’ x Récital), which provided APR to common northern Europe isolates pre-2011, has been fully defeated by more recent *Pst* isolates post-2011[67]. Conversely evidence regarding a link between the HR, a hallmark of race-specific resistance, and APR genes has been noted. APR QTL located on chromosome 2D and 4B in the cultivar Alcedo were indeed associated with a rapid and confined necrotic response similar to a HR[68]. Consequently, a gene belonging to the NLR family could actually confer these APRs. Evidence in favour of *Yr12* APR being an NLR has also been recorded (Simon Berry, personal communication). Uncovering the nature

of *Yr12* would confirm or refute whether APR strictly involves non-NLR genes or not. This question will be addressed in Chapter 2 of this thesis.

1.4. Molecular mechanisms of resistance in plants

1.4.1. The plant immune system

Traditionally, the plant immune system was described as a four-phases system[69]. The first layer of plant defences includes pathogen recognition receptors (PRRs) located in the cellular membrane that recognise pathogen-associated molecular patterns (PAMPs) that are conserved among pathogens. This first response is called PTI for PAMPs-triggered Immunity. Host-adapted pathogens can suppress PTI via secreting virulent factors called effectors within the plant cell, leading to ETS (Effector-triggered Susceptibility, second phase). Plants have evolved intracellular receptors to recognise the effectors. These intracellular receptors mostly belong to the NLR family described in section 1.3.3.1. Successful recognition of effectors by NLRs leads to ETI (Effector-triggered Immunity, third phase) that is an amplified response compared to PTI and ultimately leads to a hypersensitive response (HR) characterised by cell-death at the infection site. The fourth phase in characterised by the arms race between host and pathogen with the latter shedding/diversifying its effector repertoire and the former evolving new specificities to be able to trigger new ETI. This model was described in 2006 by Jones and Dangl and is called the 'zigzag model'[69] (Figure 1-6).

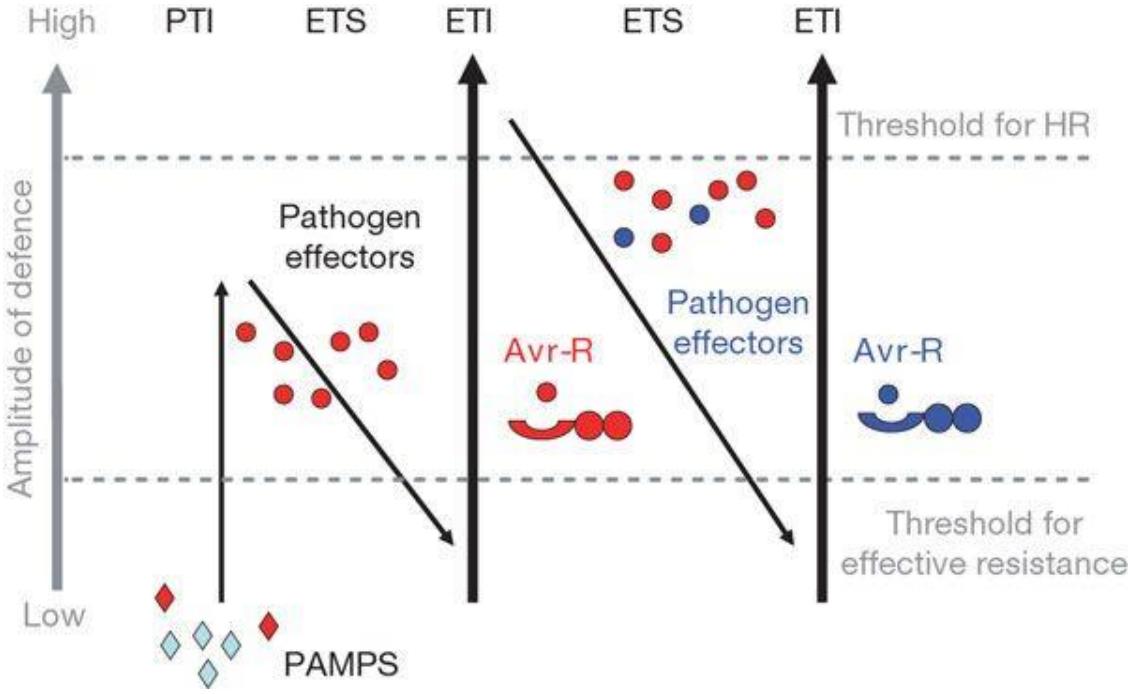


Figure 1-6. The zigzag model described by Jones and Dangl (2006)[69].
Reprinted from The plant immune system, Jones and Dangl, 2006 with permission from Springer Nature, License Number: 4700151101571.

However, the dichotomy between PTI and ETI is an on-going discussion. Among several examples reviewed by Thomma et al.,2011[70], it has been shown that certain PAMPs are specific and not necessarily widely conserved between pathogens[71]. Furthermore, van der Burgh and Joosten recently proposed to define the different forms of plant immunity solely based on the site of microbe recognition[72], where pathogens secrete intra- and extracellular immunogenic patterns (InIPs and ExIPs, respectively) that lead to activation of host response upon recognition (extracellularly and intracellularly triggered immunity (ExTI and InTI), Figure 1-7). Although the purpose of this thesis is not to discuss the best terminology to define the plant immune system, it is important to acknowledge that continuously uncovering new resistance mechanisms will help improve our understanding of disease resistance in plants and keep challenging the currently proposed models.

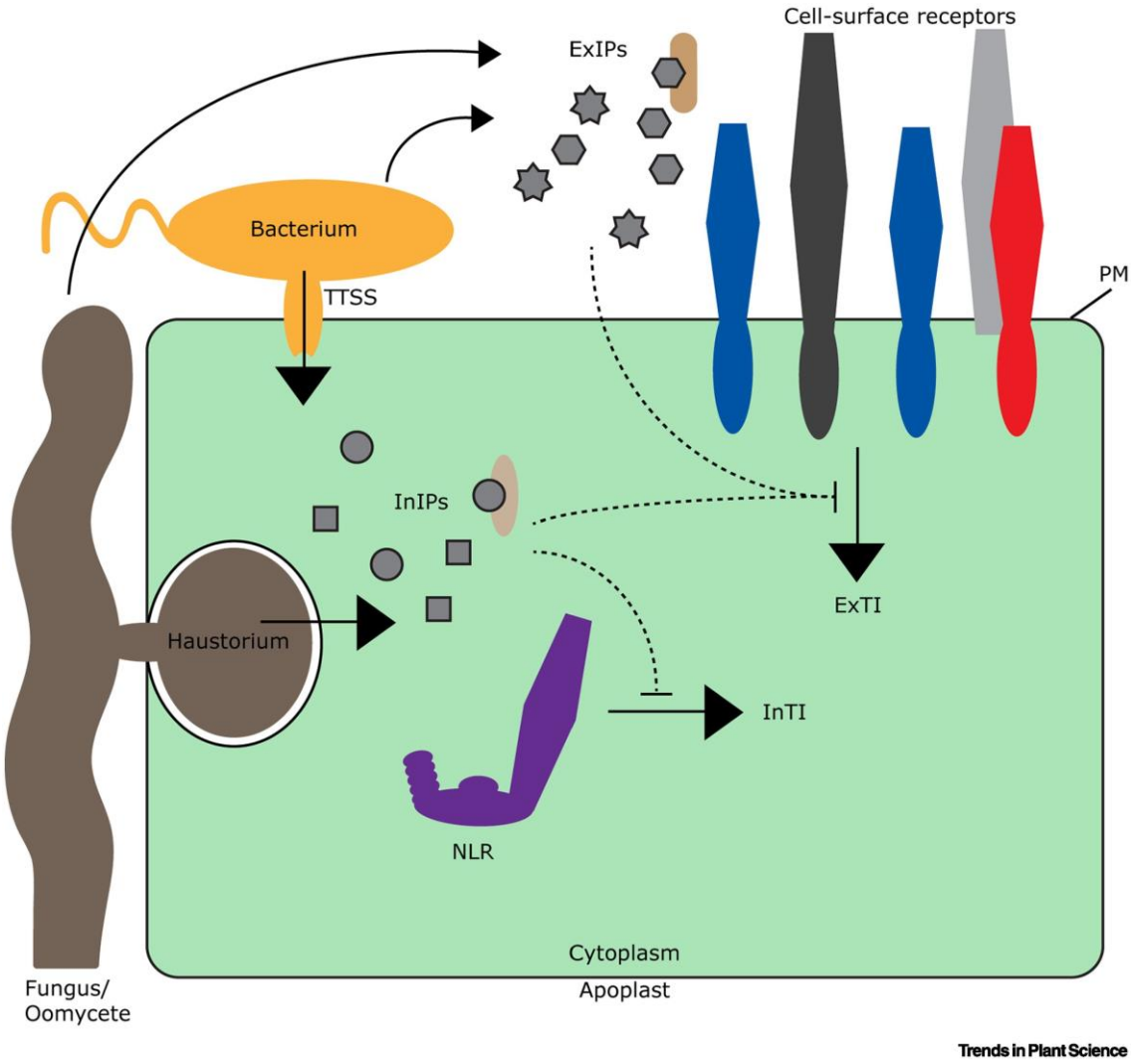


Figure 1-7. Schematic Overview of the 'Spatial Immunity Model' described by van der Burgh and Joosten, 2019[72].
Pathogens secrete intra- and extracellular immunogenic patterns (InIPs and ExIPs, respectively). Successful recognition of the ExIPs by the host's cell-surface receptors leads to extracellularly triggered immunity (ExTI), whereas intracellular recognition mediated by NLRs is called intracellularly triggered immunity (InTI).
Reprinted from Plant Immunity: Thinking Outside and Inside the Box, Vol 24, van der Burgh and Joosten, 2019 with permission from Elsevier. Licence number: 4700170542173

1.4.2. Identified molecular mechanisms involved in pathogen recognition

Kourelis and Van der Hoorn[73] recently presented an elegant meta-analysis including 314 cloned *R*-genes and describing nine different modes of actions in which these proteins trigger disease resistance. We will provide a summary of these nine mechanisms in this section and the particular mode of recognition that is relevant to this thesis will be further described in Chapter 4. As we mentioned in section 1.3.3.3, the authors used the term *R*-gene to include both PRRs and NLRs.

- Direct perception at the cell surface: This was the mechanism described as underlaying PTI in the section above. Numerous PAMPs are recognised by surface cell receptors including Receptor-like Kinases (RLKs) and Receptor-like Proteins (RLPs). The most studied PAMP in plants is bacterial flagellin, recognised by the RLK FLS2 (Flagellin-Sensitive 2)[74]. Many PAMPs are directly recognised in a similar manner to flagellin[75].

- Indirect extracellular effector perception: Modification of host proteins can also be detected outside the cell and lead to activation of defense response. *Cf-2* is tomato RLP that recognises the fungal effector *Avr2* from *Cladosporium fulvum*[76]. The recognition of *Avr2* is dependent on *Rcr3*, which encodes a secreted papain-like Cys protease[77]. Given that *Avr2* directly interacts with *Rcr3*, it has been proposed that *Cf-2* guards *Rcr3* and its interaction with *Avr2* triggers defense response. Interestingly, this example also illustrates that effector recognition does not necessarily occur inside the cell.

- Direct Intracellular Effector Recognition: This mechanism relies on NLR-mediated recognition within the cell. Numerous examples have been described in

the literature[73]. In most of the cases, the LRR region is responsible for direct binding to the effector (e.g flax L5, L6, and L7 NLRs that recognise different variants of AvrL567[78,79]). However, it has been shown that intra-protein interactions between NB-ARC and LRRs were also important in L5 and L6 and that binding of the effector competes with these interactions[80].

- Indirect Intracellular Recognition: Interactions between effectors and host proteins (either direct interaction or enzymatic modification) can also be perceived by NLRs. These host proteins have been called 'guardees' or 'decoys' depending on whether they conserved their initial activity in the plant or mimic the actual effector target[81]. For instance, ZAR1 from Arabidopsis (stands for HOPZ-ACTIVATED RESISTANCE 1) is a conserved CC-NLR that guards several class XII pseudokinases (ZED1, ZRK3, and RKS1) and the decoy kinase PBL2 (PBS1-LIKE PROTEIN 2)[82], enabling recognition of effectors derived from different pathogens, including *P. syringae* Type-III effectors HopZ1a[83] and HopF2a[84] (via ZED1, ZRK3, and RKS1) and the *Xanthomonas campestris* Type-III effector AvrAC via PBL2 and RKS1[82]. This is a remarkable example of NLR able to perceive different enzymatic activities induced by effectors derived from different pathogens. Furthermore, the structure of the ZAR1/RKS1/PBL2 complex forming upon AvrAC recognition was recently resolved showing a large pentameric active form called the 'resistosome'[85,86]. This constituted a consequent milestone in plant immunity and raised numerous questions regarding the function of NLR receptors involved in mechanisms that are different from this one.

- NLR-IDs: Many NLRs containing non-canonical domains have been identified in plants[87,88]. Three well-characterised examples showed that this 'integrated domain' is involved in direct effector recognition and led to the proposition of the 'integrated decoy model' to explain their mode of action[44]. We describe the model and its example in further details in Chapter 4, as it is relevant to the three NLRs we cloned (*Yr7*, *Yr5* and YrSP).

- Executor Genes: Executor genes have been defined as *R*-genes that are transcriptionally activated by transcription activator-like effectors (TALEs) produced by *Xanthomonas* species and confer immunity to the *Xanthomonas* strains carrying these TALEs. TALEs bind to the *cis*-regulatory elements of host targets and induce the expression of susceptibility factors. The executor gene counteract this via functioning as 'promoter-trap' for TALEs, leading to the induction of genes involved in immunity. Rice *Xa27*, which encodes a protein with multiple putative transmembrane domains, was the first executor gene being characterised[89]. Remarkably, the knowledge gained on TALEs specificity for certain DNA motifs enabled designing synthetic executor genes that provide resistance against multiple *Xanthomonas* strains[90–92]. This thus constitutes a promising strategy to engineer resistance against this type of effectors[93]. Alternatively, two recent studies successfully used CRISPR-Cas9-mediated genome editing to introduce mutations the *cis*-regulatory elements recognised by TALEs in three host sucrose transporter genes *SWEET11*, *SWEET13* and *SWEET14*, leading to resistance in an otherwise susceptible rice variety[94,95].

- Active Loss of Susceptibility: This mechanism includes host proteins that have evolved the ability to disarm the pathogen via actively altering a key process. For

example, maize *Hm1*, the first *R*-gene cloned, encodes a NADPH-dependent reductase that is specifically involved in detoxifying HC toxin produced by *Cochliobolus carbonum*[96].

- Passive Loss of Susceptibility: This mechanism involves loss of interaction between a host susceptibility factor and the pathogen effector and is common in recessive *R*-genes. For example in the case of plant potyviruses it has been shown that very specific mutations in translation initiation factors from the host prevented their interaction with the cap structure on viral transcripts and led to resistance[97].

- Passive Loss of Susceptibility by Host Reprogramming: This is typically the mechanism involved in the APR that we described in the section 1.3.3.2 (see the examples *WKS1* and *Lr34*). However, as we mentioned earlier, the exact mechanisms related to these APR genes are unknown yet. Kourelis and Van der Hoorn[73] suggested that the resistance these genes confer is seemingly dependent on a deregulated initial immune response, resulting in a quicker and stronger immune response that is able to partially suppress the pathogen.

Although more than 80 yellow rust resistant genes have been described in wheat[98], less than a handful have been cloned. We illustrated through the examples above that understanding the molecular mechanisms underlying disease resistance could help developing new strategies in breeding programs. However, because of the features we describe in section 1.5, cloning genes in wheat has been challenging. In the last decade and with the support of technological improvements, however, numerous resources have

been generated for polyploid wheat. We will demonstrate in the following section that wheat can now be used as a model crop for gene cloning.

1.5. New technologies enabled development of numerous genomic resources for wheat

In the past few years, the amount of genetic and genomic resources that have been made available for wheat dramatically increased. We contributed to a recent review giving a general overview of these resources and their potential use in functional studies in wheat[99], and to the website www.wheat-training.com that contains more detailed information on how to use these resources. We will describe in this section the resources that were relevant to this PhD.

1.5.1. Genome assemblies for wheat

Over the course of this PhD, wheat genome assemblies moved from highly fragmented and pseudo-ordered contigs to nearly fully assembled pseudomolecules (Table 1-1). Having a complete wheat genome reference sequence accelerated and facilitated gene cloning[100–102]. The large genome size (~ 15 Gb) and high proportion of repetitive elements (~ 80 %) hindered efforts to fully sequence and assemble the genome. Bread wheat is also a recent polyploid and contains three genomes: AABBDD (Figure 1-2). This determines that subgenomes carry complementary sets of homoeologous genes in collinear order across individual chromosomes, which share over 95 % sequence identity across coding regions[103]. We will describe the main wheat assemblies in chronological order of release and their characteristics in the following sub-sections:

1.5.1.1. Chromosome Survey Sequence (CSS)

The landrace Chinese Spring (CS) was chosen as the reference genome sequence given its previous use for cytogenetic studies and availability of aneuploid lines[104]. The International Wheat Genome Sequencing Consortium (IWGSC) released the Chromosome Survey Sequence (CSS) of Chinese Spring in 2014[105] (one year before the start of this work). Flow-sorting and subsequent Illumina next generation sequencing of individual chromosome arms were used to generate this assembly, allowing for separation of the three sub-genomes. Although the assembly was very fragmented, population sequencing enabled ordering contigs into genetic bins[106] (Table 1-1). However, the order of contigs within a bin was unknown and this was an issue for very wide bins. The highly fragmented nature of this assembly made difficult accurate gene annotation based on RNA-Seq data and information from related species. Nevertheless, the CSS assembly was a consequent step in wheat research and several publicly available resources kept the information about the CSS gene models as reference (including TILLING mutants and expression browser[107,108], section 1.5.3). The CSS assembly is also maintained on the archived EnsemblPlants (http://mar2016-plants.ensembl.org/Triticum_aestivum/Info/Index).

1.5.1.2. TGACv1

A more contiguous assembly of Chinese Spring was released in 2017 by The Genome Analysis Centre (TGAC) (now Earlham Institute, http://pre.plants.ensembl.org/Triticum_aestivum)[109]. The authors combined whole genome shotgun sequencing (WGS) with the newly developed W2RAP assembler[110] and achieved a higher contiguity with scaffolds around 20 times longer than is the CSS assembly (Table 1-1). These scaffolds were ordered the same way as for CSS. A new gene annotation was generated based on this assembly and gene models were overall

more complete than the CSS gene models[109]. Expression data from expression browser expVIP (http://www.wheat-expression.com/) were remapped to TGACv1 assembly. Sequencing data from TILLING mutants (section 1.5.3) were projected onto this assembly but not remapped (http://oct2017-plants.ensembl.org/Triticum_aestivum/Info/Index). The TGACv1 assembly is available on the archived EnsemblPlants. An even more contiguous assembly combining short Illumina reads and very long Pacific Biosciences reads was released the same year[111] (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA392179). However, no gene models were released with this assembly.

1.5.1.3. IWGSC RefSeqv1.0

The IWGSC released last year the first wheat assembly with most of the contigs organised into 21 pseudomolecules[112] (Table 1-1). To achieve this, the consortium generated Illumina sequencing data and used the assembler DeNovoMAGIC (https://www.nrgene.com/solutions/denovomagic/). The addition of Hi-C data and population sequencing enabled the assembly of pseudomolecules. Numerous data have been compiled to further improve the assembly and generate what is now RefSeqv1.0 (described here https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies). A new gene annotation was performed on RefSeqv1.0, which is publicly available (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations). There is now an updated annotation including manually corrected genes from RefSeqv1.0 (RefSeqv1.1). Earlier this year, RefSeqv2.0 was released and it included optical maps to enable chimeric scaffold identification and correction. Additionally, the consortium also compiled the PacBio data generated in 2017[113] to perform gap closing in RefSeqv1.0 and generate RefSeqv2.0. However, this new assembly was released too recently to allow its inclusion in this thesis. For this work, we thus used RefSeqv1.0 and both its annotations

(RefSeqv1.0 for the Synteny analysis and RefSeqv1.1 for the Neighbour-net analysis in Chapter 4). This version is also the one currently available on EnsemblPlants (https://plants.ensembl.org/Triticum_aestivum/Info/Index) and both expression data from expVIP (www.wheat-expression.com) and TILLING mutant sequencing data (www.wheat-tilling.com, section 1.5.3) from were remapped to RefSeqv1.0 and RefSeqv1.1 gene models. Additionally, a new expression browser was developed to illustrate and quantify expression data from a developmental time-course in the variety Azhurnaya (http://bar.utoronto.ca/efp_wheat/cgi-bin/efpWeb.cgi, eFP).

Table 1-1. Summary of the statistics of three selected wheat assemblies that were relevant to the work presented in this thesis. Adapted from Adamski et al., 2018[99].

	CSS	TGACv1	RefSeqv1.0
Release date	IWGSC, 2014	Clavijo et al., 2017	IWGSC, 2018
# contigs/ chromosomes	> 1 million	735,943	21 chromosomes + Un
Mean scaffold size	7.7 kb	88.7 kb	Chromosomes
Assembly Size	10.2 Gb	13.4 Gb	14.6 Gb
Order	Crude order	Large Bins	"True" physical order
# Coding genes	100,934	104,390	107,891 High Confidence 161,537 Low Confidence
Resources	Archive EnsemblPlants	Archive EnsemblPlants	EnsemblPlants*
	TILLING mutants	TILLING mutants	TILLING mutants
	expVIP	expVIP	expVIP + eFP
Accession	Chinese Spring	Chinese Spring	Chinese Spring

Archive EnsemblPlants CCS: http://mar2016-plants.ensembl.org/Triticum_aestivum/Info/Index
Archive EnsemblPlants TGACv1: http://oct2017-plants.ensembl.org/Triticum_aestivum/Info/Index
EnsemblPlants RefSeqv1.0: https://plants.ensembl.org/Triticum_aestivum/Info/Index
*Includes SNP variation, gene trees, homoeolog assignments

1.5.1.4. Genome assemblies of wild progenitors of wheat

In this thesis we also used pseudomolecule-assembled genomes of the D genome donor of hexaploid wheat *Aegilops tauschii* and the corresponding gene annotation[114] in the synteny analyses presented in Chapter 4. We included the recently sequenced wild emmer Zavitan and its gene models[23]. It is important to acknowledge that assemblies for *Triticum urartu*[115] and domesticated emmer Svevo[116] were also released in the past years, although we did not include them in this work.

A remaining challenge is the fact that there might be presence/absence variation in gene content between the Chinese Spring reference and wheat varieties of interest. Consequently, mapping these genes onto the reference might not be possible. For resequencing of varieties, it will thus be important to take advantage of the un-mapped reads (against the standard Chinese Spring reference) to ensure that information is not lost. Alternatively, several high-contiguity genome assemblies for elite cultivars have also been released during this thesis and we will describe them in the following section.

1.5.2. Towards a pangenome of wheat

The pangenome represents the entire gene set of all varieties of a species. It includes genes present in all strains (core genome) and genes present only in some strains of a species (variable or accessory genome). Certain gene families have important presence/absence polymorphisms across strains of a species (e.g. resistance genes). It is thus valuable to have as much information as possible from different strains to have a good representation of the whole gene family. Sequencing several varieties and generating reference-quality assemblies for each remains costly for species with large and complex genomes such as hexaploid wheat.

During this PhD, 14 wheat assemblies were released as part of the global '10+ Wheat Genomes Project' (http://www.10wheatgenomes.com). Four UK bread wheat cultivars (Cadenza, Paragon, Claire and Robigus) and one pasta wheat cultivar Kronos were sequenced and assembled by the Earlham Institute the same way as described above for TGACv1 (Chinese Spring). These assemblies were released in 2017 and are available for download and BLAST analyses (https://wheatis.tgac.ac.uk/grassroots-portal/blast). An additional set of nine international varieties including CDC Landmark and CDC Stanley from Canada, Mace and Lancer from Australia, Jagger from the USA, Julius from Germany, Arina*LrFor* from Switzerland, SY-Mattis from France and Norin61 from Japan were sequenced and assembled in a similar manner to RefSeqv1.0 of Chinese Spring. These resources were critical for Chapter 3 and 4 and we provide more information in the dedicated sections.

1.5.3. Exome-sequenced mutant populations

We mentioned in section 1.5.1 two exome captured and sequenced TILLING mutant populations whose sequencing data have been mapped to the different Chinese Spring gene models. This resource was published in 2017[107], however, the mutant lines and associated sequencing data were already available at the start of this PhD in 2015 (https://www.seedstor.ac.uk/search-browseaccessions.php?idCollection=24 for the mutant lines and http://www.wheat-tilling.com/ for sequencing information).

These populations were generated in the UK hexaploid cultivar Cadenza and the tetraploid cultivar Kronos, which were also sequenced and assembled recently (section 1.5.2). This is a highly valuable resource for reverse genetics, as it allows the identification of lines carrying specific mutations in the gene(s) of interest and thus further investigation of a possible associated phenotype. The database includes exome

sequences of 1,535 Kronos and 1,200 Cadenza mutants that have been re-sequenced using Illumina next-generation sequencing. Mutations were identified, and their effects predicted based on the protein annotation available at the Ensembl Plants website for IWGSC gene models (https://plants.ensembl.org/Triticum_aestivum/Info/Annotation/) and www.wheat-tilling.com for CSS gene models. For this project, we used the Cadenza mutant population to clone one of the targeted *Yr* genes (*Yr7*) and we provide more details in Chapter 2.

These new resources marked a major turn in wheat research. Indeed, functional studies on agronomically important traits can now be performed in wheat directly without having to systematically rely on other model plant species. We only described here the resources that we exploited in the frame of this PhD and further details are available on www.wheat-training.com and in the recent review by Adamski et al., 2018[99]. One of the main focus of this thesis was studying disease resistance in wheat via identifying yellow rust resistance genes. We will thus discuss how this can be done in wheat in the following section.

1.6. Studying traits in wheat with forward genetics

Forward genetics is the approach of determining the genetic basis responsible for a phenotype. In other words, no prior knowledge is known about the nature of the genetic variant(s) that are involved in the expression of the trait of interest. This can be done by using naturally occurring or induced mutations. Most of the techniques used in this thesis rely on forward genetics and we will thus describe the main approaches that have been developed for this purpose in wheat and other plants. We also will provide a brief description of reverse genetics approaches in wheat to illustrate what can now be achieved in this crop.

1.6.1. Map-based cloning

1.6.1.1. Principle:

The critical step of candidate gene identification by mapping is the ability to rapidly fine-map the phenotype to a very narrow genetic interval. Genetic mapping relies on two major processes. Recombination frequency represents the number of crossovers occurring between two loci and thus assesses their genetic linkage given that the closer two loci are to each other, the less likely a crossover event will occur and vice-versa. Genetic mapping thus relies on the development of large populations to increase the probability to obtain recombinants.

Recombination alone is not sufficient to draw a genetic map if nothing could differentiate the parents in terms of sequence variations (namely markers). Polymorphism is thus also crucial for genetic mapping and correspond to the basic blocks constituting a genetic map. Molecular markers are specific fragments of the genome sequence. There are several types of molecular markers, including Restriction Fragment Length Polymorphisms (RFLPs) that correspond to restriction sites and Simple Sequence Repeats (SSRs), Sequence Tagged Sites (STSs) and Amplified Fragment Length Polymorphisms (AFLPs) that are PCR-based markers. We will discuss in section 1.6.2 Single Nucleotide Polymorphism markers. Differences in the DNA sequence of the parents are used to generate a genetic map that illustrates the genetic linkage between these markers. The genetic map thus needs to be dense enough to cover most of the genome and once the trait of interest has been defined within flanking markers, one can subsequently saturate the interval with additional markers.

1.6.1.2. Map-based cloning in bi-parental populations:

Map-based cloning in bi-parental populations relies on using two parents that are different for the phenotype of interest (e.g. resistance or susceptibility to a pathogen strain) to generate a population that segregates for this trait. Before high-contiguity assemblies were generated for wheat, the only accessible wheat sequences were randomly cloned into bacterial artificial chromosomes (BACs). These fragments were small and corresponded to 100-200 kb of DNA each. This represents 0.001 % of the wheat genome and it thus required ~ 500, 000 BAC clones to cover the whole genome[117]. Because of their small size, it was crucial to narrow down the genetic interval (and thus physical interval) as much as possible to encompass only few overlapping BACs. Hence the necessity to develop large mapping populations to increase the chance of obtaining recombinants. However, there was still the issue of low-recombination rate regions (e.g. centromeric regions), which are common in wheat[112]. Nevertheless, genetic mapping was successful in cloning genes in wheat, especially disease resistance genes (reviewed in Keller et al., 2018[100]).

1.6.1.3. QTL mapping in Near Isogenic Lines:

Many agronomically important traits rely on several genes for the expression of the related phenotype and are quantitative (Quantitative Trait Loci, QTL). However, in most cases there are single genes that account for a significant part of the phenotypic variance. Map-based cloning can thus be used to identify such targets. Development of Near Isogenic Lines is frequent in this approach as it allows introgressing the trait of interest in the background of a line that does not show the phenotype of interest. Thus, all the loci that do not contribute and are not genetically linked to the trait are segregated away. The first QTL in wheat was isolated via map-based cloning[118].

1.6.2. Association genetics

Genome-wide association study is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is statistically associated with a trait. GWAS typically focuses on Single Nucleotide Polymorphisms (SNPs), although other markers can also be used (e.g. *k*-mers, see below). With the first reference genome released with its gene annotation[105], SNPs could now be accessed in wheat via re-sequencing varieties and sequence comparison to the reference genome. SNPs- based markers are widely used nowadays and their advantages have been reviewed elsewhere[119]. There are abundant SNPs between varieties so these markers can generate fairly dense genetic maps. SNPs can also be assayed using high-throughput genotyping methods, such as Kompetitive allele specific PCR (KASP).

Numerous SNP arrays have been developed for wheat. They allowed assaying of genetic diversity within wheat and most SNP arrays were created based on coding sequence and UTRs plymorphisms[120–122]. They provided a common base for comparison of thousands of varieties and landraces and corresponding datasets are available online (e.g CerealsDB, https://www.cerealsdb.uk.net). This constitutes valuable resources for both fundamental and applied research.

During this PhD, a new technique combining targeted sequencing with association genetics was successfully used to clone resistance genes from *Aegilops tauschii*[123]. Instead of investigating SNPs among their diversity panel, the authors explored the association between *k*-mers (DNA sequences that are *k* bp long) and the resistant/susceptible phenotype displayed by the different lines. This is very powerful, especially to identify variation that cannot be accessed via mapping reads to a reference genome (e.g. complete absence of the locus of interest in the reference genome). The

authors combined this association genetics approach with resistance gene enrichment sequencing (RenSeq, see section 1.6.4.2) to 're-clone' three resistance genes originating from *Aegilops tauschii* (*Sr33* and *Sr45*) and clone *Sr46* and *SrTA1662*[123].

1.6.3. Mutant screens

We illustrated in the section above how natural variation could be accessed in wheat to identify specific genetic variants linked to a trait of interest. Alternatively, induced variation is also a powerful tool to achieve similar goals. Mutant screens rely on the development of a large population of chemically (via application of chemical compounds such as ethyl methane sulfonate, sodium azide) or irradiated (X-ray, gamma-rays, UV light, fast neutrons) mutagenized individuals from a well-characterised genetic background. Each method can give rise to a wide range of different types of mutations such as single base substitutions, insertions/deletions, duplications. In wheat, seeds are the preferred tissue for mutagenesis treatment as it is a sexually propagated crop and the aim is to generate heritable mutations for further studies in the descendance and crosses. The mutants are screened for the phenotype of interest and genetic mapping is performed to locate the causal mutation. The way of identifying the responsible gene(s) evolved over the years thanks to various technological and methodological improvements[124].

1.6.4. Mapping-by-sequencing

Numerous next-generation sequencing (NGS) mapping approaches have been developed in diverse organisms. A summary of the main approaches commonly used in plants is reported in Table 1-2 to allow their comparison. Two main characteristics define them: whether they are based on whole genome resequencing or reduced representation sequencing, and whether they rely on recombination or not (Table 1-2). Mapping by

sequencing methods are based on Bulked Segregant Analyses (BSA)[125] and SNPs filtering and frequency analyses. BSA facilitates the linkage assessment between the phenotype of interest and a wide range of genetic markers via considering the allele frequencies in a population of recombinants with similar phenotypes, usually called a bulk. It enables phenotype association to a genomic region instead of genotyping every individual one by one[125].

Varietal or mutational SNPs may be evaluated to map the causal mutation. An unlinked SNP would segregate randomly in the progeny with a frequency close to 0.5 in both mutant and wild-type reads, whereas a linked SNP would segregate with the phenotype. Its frequency would thus be approaching 1.0 in the mutant reads and 0 in the wild-type reads (Figure 1-8).

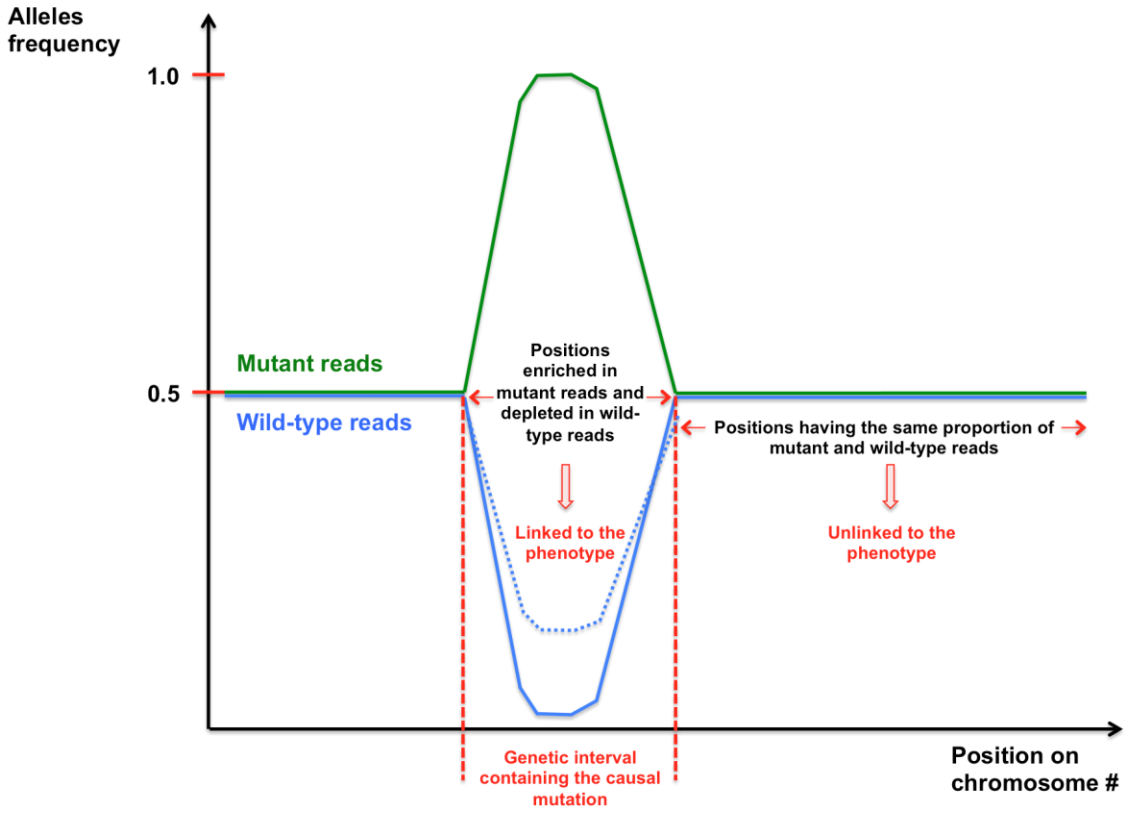


Figure 1-8. Schematic representation of the output of a mapping by sequencing techniques.

The alleles frequencies are plotted across one chromosome (it could also be a contig or a scaffold). The mutant reads are in green and the wild type ones are in blue. An unlinked marker would segregate randomly within the progeny and would thus be equally represented in the wild-type and mutant reads. A linked marker, however, would be present only in the mutant reads. This leads to an increase of its frequency in the mutant reads and a decrease in the wild-type ones. The genetic interval is the chromosomal region linked to these variations of the marker frequencies. The blue dash line shows the case when heterozygous and homozygous wild-types individuals could not be differentiated. This leads to a lighter depletion of reads carrying the linked marker in the wild-type bulk, as the heterozygous individuals carry both the unlinked and linked marker.

1.6.4.1. WGS-based techniques

An approach called SHOREmap successfully identified a causative mutation in Arabidopsis by Illumina sequencing[126]. This two-step analysis first identifies the genetic interval that likely includes the causal mutation by analysing marker frequency in sliding windows of 200kb in a similar way to the one presented in Figure 1-8. Then the markers are ranked according to their distance to the marker distribution. If a gene annotation is provided, the technique can predict the effect of the mutation on the protein model. The authors were able to fine-map and clone the gene of interest while studying the same population, which was very impressive. A lot of recombinants (500 F_2 plants were pooled) were needed to reach that accuracy. Similar methods have been developed afterwards: Next Generation Mapping (NGM)[127] allows reducing the number of requested recombinants (10 to 80 F_2 lines were pooled) and MutMap permits using the same cultivar for the F_1 cross, as it assesses SNPs incorporated by the mutagenesis (EMS in the study) as markers[128]. An improved version of MutMap, namely MutMap+[129], now allows causal mutation identification within the mutant lines themselves without requiring any backcrosses to the wild-type. This is advantageous especially when recombination suppression can be observed in the targeted region.

1.6.4.2. Reduced representation sequencing-based techniques

Large and high repetitive genomes still made WGS analyse challenging. When the subsequent analyses aim to focus only on gene mutation, one could consider only sequencing the coding part of the genome to reduce its complexity. Combination of RNA-Seq and bulked segregant analysis enables gene fine-mapping in tetraploid wheat and maize[130,131]. This approach has been successfully used to identify high-resolution genetic markers for breeding in hexaploid wheat[132]. The strategy is similar to the one used with WGS-based studies (Figure 1-8). Sequencing the exomes, which represents

the coding region of the genome, can address the expression variation issues underlying RNA-Seq studies. A combination of TILLING and exome capture and sequencing has been successfully applied to hexaploid wheat for mutation detection[133]. The reference assembly quality is crucial for the reduced representation sequencing-based methods, as they rely on mapping the reads onto the reference for SNPs calling.

Alternatively, providing the information about the chromosome location of the gene of interest is known, isolating this specific chromosome for targeted sequencing has proven to be a successful technique for cloning genes in wheat[134]. The resistance gene *Pm2* was cloned using this technique. The authors combined flow-sorting and short read sequencing of chromosome 5D from variety carrying *Pm2* and six independent EMS-mutagenised *Pm2* loss of function mutant to identify the causal mutation and the gene of interest. This approach is called MutChromSeq and is useful when the reference genome(s) does not contain the gene of interest. Another technique relying on chromosome flow sorting and long-range scaffolding could also be an alternative to clone genes absent from the reference genome(s)[135]. This technique requires prior map-based cloning to identify the smallest genetic interval possible that contains the gene of interest. After that, the corresponding physical interval is defined via a combination of short-read sequencing with chromosome contact maps of chromosomes reconstituted in vitro (Dovetail Genomics Chicago method[136]). This approach was successfully used to clone the resistance gene *Lr22a* in wheat[135].

Targeted sequencing can also be applied to specific gene family displaying a characteristic domain pattern that is unique to this family. For example, numerous cloned resistance genes belong to the NLR family (Nucleotide-Binding Leucine Rich Repeat, described in section 1.3.3.1), which correspond to this criterion because of their specific

domain organization. This pattern has been successfully used for targeted enrichment sequencing in potato and tomato[51,52]. BSA combined with RenSeq enables SNP calling within the F_1 progeny from two different potato species to produce markers closely linked to the targeted resistance gene[52]. More recently, a pipeline named MutRenSeq was developed in wheat to increase RenSeq accuracy and allow the identification of the targeted NLR gene itself[137]. This three-step method is based on a typical EMS-mutagenesis screen for susceptible mutants, followed by a RenSeq on the independent M_2 lines and on the non-mutagenized parental line and sequence comparison via a presence/absence SNP calling. MutRenSeq enabled the cloning of two additional stem rust resistance gene in wheat (*Sr22* and *Sr45*). We used this method to clone *Yr7*, *Yr5* and *YrSP* and provide further technical details in Chapter 2 and Chapter 3.

It is important to notice that the "best method" does not exist, each of them has its pros and cons and the aim is to select the one that best suits the model of study. For example, Arabidopsis has a short generation time, thus methods relying on recombinants are still relevant whereas it could be more problematic for wheat, for which direct mutant sequencing is more convenient. The significant progresses that have been made in terms of genome assembly in hexaploid wheat along with the implementation of new methods to analyse the data provided by the NGS technologies are very promising regarding gene fine-mapping and cloning. Indeed, it enables a better ability to deal with this large and repetitive genome.

Table 1-2. Table comparison of the mapping by sequencing approaches developed in plants

	Technique (species)	Crossing scheme	Recombinants needed?	Pros	Cons	References
Whole genome re-sequencing	**SHORE-map** (*Arabidopsis thaliana*)	**Classical mapping population** Outcrossing the mutant to a genetically diverges line followed by one self-cross	Yes Bulked Segregant Analysis	- Gene of interest can be found in one go - Mutation identification despite mutant crossed to a diverged strain	- Screening of 500 EMS-derived F_2 lines - Difficult to apply to dominant mutations - Expensive for large genomes	Schneeberger et al., 2009
	NGM (*Arabidopsis thaliana*)	**Classical mapping population** Outcrossing the mutant to a genetically diverges line followed by one self-cross	Yes Bulked Segregant Analysis	- Gene of interest can be found in one go - Mutation identification despite mutant crossed to a diverged strain	- Screening of 10-80 EMS-derived F_2 lines - Difficult to apply to dominant mutations - Expensive for large genomes	Austin et al., 2011
	MutMap (*Oryza sativa*)	**Isogenic mapping population** Crossing homozygous mutants to the non-mutagenised parent	Yes Bulked Segregant Analysis	- F1 cross involves the wild-type parent and the mutant from the same cultivar - Only one F_2 line is required	- Markers need to be identified by a *de novo* search for segregation in the pool - Resolution might not be good enough to clone the targeted gene in one go	Abe et al., 2012
	MutMap+ (*Oryza sativa*)	**Homogeneity mapping** Exclusively analyzing the genomes of the affected individuals	No	- Does not rely on recombination - Resolution can be improved by sequencing a pool of non-mutant siblings	- Markers need to be identified by directly sequencing the mutant genome	Fekih et al., 2013
Reduced representation sequencing	**RNA-seq** (*Triticum aestivum*, *Zea mays*)	**Isogenic mapping population** Crossing near isogenic lines followed by a self-cross	Yes Bulked Segregant Analysis	- Focuses on the coding regions of the genome to target gene-specific mutations - Cost effective - Differential expression analyses can be carried out in parallel	- Differences in gene expression and allele-specific expression add another source of variation	Trick et al., 2012; Lu et al., 2012; Ramirez-Gonzalez et al., 2015
	Exome capture (*Triticum aestivum*)	**Direct targeted sequencing of independent mutant genomes** **Or** **Isogenic mapping population** Crossing homozygous mutants to the non-mutagenised parent	No Yes, Bulked Segregant Analysis	- Focuses on the coding region of the genome to target gene-specific mutations - Allows direct mutation identification, providing that the reference sequence quality is high enough	- Causal mutations that are not located within the coding region itself but in a regulatory element will not be considered - Relies on capture with pre-designed baits so only capture what is already known	King et al., 2015
	MutChromSeq (*Triticum aestivum, Hordeum vulgare*)	**Targeted sequencing of chromosomes derived from independent mutant lines**	No	- Does not rely on capture so no prior knowledge on the nature of the targeted gene needed	- Still expensive for wheat - Requires knowledge about the location of the targeted gene	Sanchez-Martin et al., 2016
	TACCA *(Triticum aestivum)*	**Targeted sequencing of the chromosome derived from the variety carrying the gene of interest**	Yes, Bulked Segregant Analysis	- Same as for MutChromSeq	- Prior fine-mapping of the gene of interest is required - Assembly technique and chromosome flow sorting is expensive	Thind et al., 2017
	RenSeq (*Solanum tuberosum*, *Solanum lycopersicum*)	**Classical mapping population** Crossing a susceptible variety with a resistant one, followed by one self-cross	Yes, Bulked Segregant Analysis	- Reduces drastically the complexity of the genome by focusing on one single gene family = Cost effective - Identification and annotation of the plant NLRs	- Enrichment is PCR-based so it can induce biases (PCR duplicated) - Relies on capture with pre-designed baits so only capture what is already known	Jupe et al., 2013; Andolfo et al., 2014
	MutRenSeq (*Triticum aestivum*)	**Direct targeted sequencing of independent mutant lines**	No	- Same as for RenSeq - Do not rely on recombination - Reference-free - Gene of interest can be found in one go	- Same as for RenSeq - *De novo* assembly of NLRs can be highly fragmented and further work may be required to identify the full length of the candidate gene	Steuernagel et al., 2016
	AgRenSeq (*Aegilops tauschii*)	**Diversity panel displaying sufficient variation in the phenotype of interest**	No	- Same as for MutRenSeq - Not mutant population required	- Same as for MutRenSeq - Targeted gene(s) need to be distributed across the breadth of diversity within the panel to increase signal/noise ratio	Arora et al., 2019

1.6.5. Note on reverse genetics in wheat

Reverse genetics refers to elucidating the function of a gene by analysing the phenotypic effects of specific alteration in its sequence. We described in section 1.5.3 the development of two exome-captured and sequenced TILLING populations in hexaploid wheat Cadenza and tetraploid wheat Kronos[107]. Although TILLING is usually a forward genetics approach, sequencing the exome of each mutant line allows the identification of all its mutations in the gene space. Consequently, it enables the selection of mutant lines carrying mutations in specific genes to study their phenotype and it is thus a reverse genetic approach. In Chapter 3, we illustrate how we took advantage of this to select mutations located within a physical interval on chromosome 2B and converted them into markers to fine-map the *Yr7* causal mutations in a Cadenza wild-type x Cadenza mutant cross.

Other reverse genetics approaches in wheat now include transgenic-based approaches to achieve transient or stable transformation. Both bombardment[138] and *Agrobacterium*-mediated transformation[139] can be performed in wheat. This allows a wide-range of available experiments including overexpression of one particular gene/allele in a wheat background that initially lacked it (e.g. expression of a resistance gene in a susceptible cultivar), reviewed in Hensel et al., 2011[140]. CRISPR-Cas9-mediated alterations of the sequence of the targeted gene also have the advantage of potentially targeting all homoeologs at once and thus constitutes one option to overcome functional redundancy across the three genomes[141]. RNA-interference (RNAi) was also successfully used in wheat to reduce gene expression in all homoeologs simultaneously[142].

1.6.6. Selecting for traits in wheat breeding

Identifying genes/alleles responsible for a given phenotype constitutes the first step in understanding what are the mechanisms underlying the expression of this phenotype. While this is obviously relevant to fundamental studies, understanding the genetic determinants of agronomically important traits is crucial to enable development of better performing varieties in the field. Originally, selection was performed at the phenotypic level only. Growers would select the plants displaying the most favourable traits for cultivation and inter-cross them to develop better varieties. Although numerous breeding programs still mostly rely on phenotype selection, it is now possible to predict the phenotype of an individual to a certain extent based on its DNA sequence.

In the section above, we discussed how genetic mapping allows identification of markers that are linked to a specific phenotype. This information is valuable to predict the phenotype of a given plant based on the presence/absence of these markers via sampling DNA from the individual at early developmental stage. Therefore, it is not required to wait until the plant reaches the developmental stage when the phenotype is expressed to select it. This is useful in breeding programs because it allows for selection of individuals to be taken to the next step based on their genotype. For example, identifying a marker 100 % linked to a resistance gene would predict that the plant will be resistant against a certain strain of a pathogen without having to challenge it. This is called Marker-Assisted Selection (MAS) and we will discuss this further in Chapter 3.

MAS is not the sole approach used in breeding to select for agronomically important traits. Indeed, it is difficult to select for rare QTLs that have a small effect on agronomically important traits but can be highly advantageous in a certain environment. A recently implemented method called Genomic Selection (GS) combines molecular and

phenotypic data in a training population to estimate the breeding values of individuals in a testing population that have been genotyped but not phenotyped (reviewed in Crossa et al., 2017[143]). Because the phenotyping step is skipped, it reduces the cost and the time necessary to develop a new variety. However, genotyping is still necessary and the populations in the training set have to be large to increase statistical power, which increases the cost. Several statistical models have been developed and it can be difficult to predict which can be efficiently applied[143]. Nevertheless, a growing number of studies demonstrated that GS could be successfully implemented in breeding programs for disease resistance[144,145].

1.7. Summary

Wheat is a globally important crop and increasing wheat yield is one of the numerous components needed to achieve food security (section 1.1 and 1.2). However, yellow rust disease is among the major biotic constraints threatening wheat yield (1.3). More than 80 yellow rust resistant genes have been described in wheat[98], whereas less than a handful has been cloned. With the technological advances made in genomics and the growing number of resources supporting functional studies in wheat (section 1.5), it is now possible to address this issue directly in wheat. Cloning resistance genes is crucial for specific marker development to assist breeding (section 1.6) and characterising molecular mechanisms involved in disease resistance (section 1.4). In this thesis, we will focus on seedling resistance conferred by *Yr7, Yr5* and *YrSP* and 'adult plant resistance' conferred by *Yr12*. We will provide a detailed introduction in Chapter 2 regarding each of these genes.

1.8. Thesis aims

This thesis aims to understand the molecular components of yellow rust resistance in hexaploid wheat *Triticum aestivum* via focusing on specific resistance genes (*Yr7*, *Yr5*, *YrSP* and *Yr12*). To achieve this aim, we will first use forward genetics to identify *Yr7*, *Yr5*, *YrSP* and *Yr12*-loss of function mutants (Chapter 2) and carry out a MutRenSeq experiment to clone the corresponding genes (Chapter 3). We will then combine comparative genomics and neighbour-net approaches to generate hypotheses regarding their mode of action (Chapter 4) and test these hypotheses (Chapter 5).

2. Forward genetic screens to identify loss of function mutants for *Yr7*, *Yr5*, *YrSP* and *Yr12*

2.1. Introduction

1.1.1. Yellow rust resistance genes investigated in this thesis

We mentioned in Chapter 1 that yellow rust disease is among the major wheat biotic constraints. Only two *Yr* genes have been cloned so far: the race-specific *Yr15*[146] which encodes a tandem kinase pseudokinase was cloned during the time of this PhD, and the adult plant resistance gene *Yr36* (*WKS1*)[63], which encodes a protein kinase with a lipid-binding domain. There is some controversy about *Yr10*: previous work demonstrated that *Yr10* was successfully cloned but a more recent study showed that the identified gene might not be *Yr10* after all[147,148]. It is also important to note that the well-studied *Lr34* encodes an ABC-type transporter and, despite its denomination, confers partial resistance to yellow rust (*Yr18*), powdery mildew (*Pm8*) and stem rust (*Sr57*). In addition, leaf tip necrosis (*Ltn1*) is a phenotypic characteristic of *Lr34*-mediated resistance[149]. Similarly, *Lr67* shows broad-spectrum, partial resistance to rust and mildew pathogens (*Lr67*/*Yr46*/*Sr56*/*Pm39*/*Ltn3*)[150] and encodes a hexose transporter[151]. Hence at the start of this PhD there were no canonical NLR genes encoding for yellow rust resistance cloned.

2.1.1.1. Origin of Yr7, Yr5 and YrSP

Yr7:

Yr7 originates from durum wheat Iumillo and has been introgressed in hexaploid wheat Thatcher, which is the main known *Yr7*-source in breeding programs. *Yr7* has been widely deployed in UK breeding programs and has been defeated for almost a decade in the field (see section below).

Yr5:

Yr5 originates from spelt wheat Album. As opposed to *Yr7*, it is still effective in the field and confer resistance to all tested *Pst* isolates but two (Table 2-1).

YrSP:

YrSP was identified in bread wheat Spaldings Prolific and still confer resistance against *Pst* races in certain regions of the world, including North America. However, *YrSP* has been defeated in the UK despite never really being deployed in commercial cultivars.

2.1.1.2. Rapid breakdown of single-deployed resistance genes: the examples of Yr7 and Yr17

Race-specific resistance genes are easily overcome when deployed on their own. One single mutation in the corresponding *Avr* gene can lead to loss of recognition and therefore loss of resistance in the host (discussed in Chapter 1). Thus, there is a very strong selection pressure on virulent pathogen strains, which consequently may overcome the *R*-gene within a few years after its first deployment. A well-known example of *Yr* breakdown is *Yr17*[152]. Between 1993 and 1997 the proportion of cultivated *Yr17* varieties in the UK increased considerably: the percentage of wheat acreage dedicated to *Yr17* varieties rose from 1 to 35 %. In parallel, the percentage of tested *Pst* isolates virulent to *Yr17* also rose from 1% in 1994 to nearly 100% in 1997. The same trend was observed in Denmark and France[152].

We observed a similar pattern for *Yr7*. We combined data from NIAB-TAG Seedstats journal (NIAB-TAG Network) to estimate the percentage of harvested *Yr7* varieties between 1996 and 2016 and from the UK Cereal Pathogen Virulence Survey

(https://ahdb.org.uk/ukcpvs) to show the evolution of the prevalence of *Pst* isolates virulent to *Yr7* over the same period of time (Figure 2-1). Although no causal link can be drawn from this graph, there is a correlation between the increase of cultivated *Yr7* varieties (green) and the prevalence of *Pst* isolates virulent to this gene (orange), consistent with the *Yr17* example mentioned above.

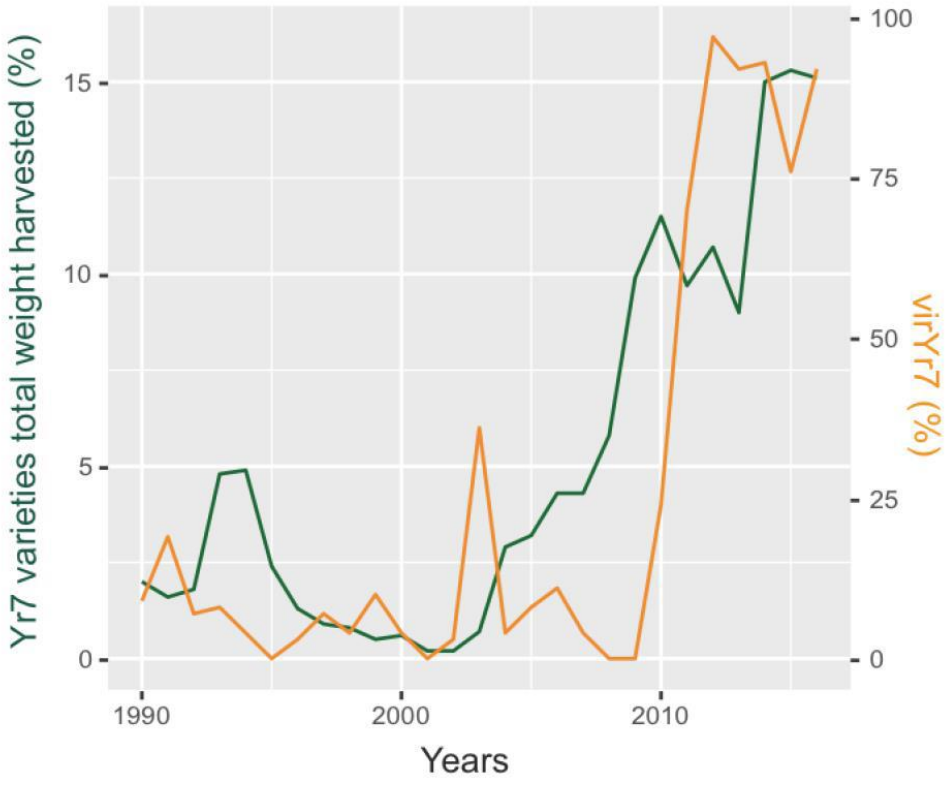


Figure 2-1. Percentage of total harvested weight of wheat cultivars carrying *Yr7* (green) and the proportion of tested *Pst* isolates virulent to *Yr7* (orange) from 1990 to 2016 in the United Kingdom
Published in Marchal et al., 2018[153], data are presented in Appendix 8-1.

Interestingly, *Yr7* has been hypothesized to be allelic to *Yr5*[154] and the latter is still effective in the field with only two of > 6,000 tested *Pst* isolates worldwide being virulent to *Yr5* (Table 2-1). It is important to note that *Yr5* has not been widely deployed on its own in the field and thus it does not mean that it drives broad-spectrum resistance. Both *Yr17* and *Yr7* examples above illustrate how quickly resistance driven by single-dominant genes can be overcome. Hence, it highlights the importance of stewardship plans to deploy *Yr5* in combination with other genes as currently done in the USA (e.g. *Yr5*+*Yr15*; UC Davis breeding programme) to avoid a *Yr7* and *Yr17*-like scenario.

Table 2-1. Summary of Pst isolates tested on *Yr5* differential lines from 2004 to 2017 in different regions.
Overall, >6,000 isolates from 44 countries displaying >200 different pathotypes were tested on *Yr5* materials and no virulence was recorded apart from two isolates from Australia in 1984, PST 360 E137 A-/+. Data were obtained from public databases and reports on yellow rust surveillance, whose references are recorded. It is important to note that we report here the number of identified pathotypes for a given region and database. Similar pathotypes could thus have been counted twice if identified in different regions

Region	Countries	Samples	Pathotypes[1]	Year(s)	Reference
Europe	22	1839	15	2009-2017	http://wheatrust.org
Africa and West/Central Asia	19	436	20	2009-2017	http://wheatrust.org
USA	1	3596	140	2004-2015	http://striperust.wsu.edu/races/data/
West China	1	308	56	2013	Zhan et al., 2016[155]
Australia	1	-	16	2005-2015	http://rustbust.com.au[156]

[1]Number of identified pathotypes for a given region and database. The same pathotype could thus have been counted twice if identified in two different regions
[2]PST 360 E137A+ and 360 E137A- are the only isolates reported to be virulent to Yr5 to date and are avirulent to *YrSP*[157]

The genetic relationship between *Yr5* and *Yr7* has been debated for almost 45 years[158,159]. Both genes map to chromosome arm 2BL in hexaploid wheat and are closely linked with *YrSP*[54] (522 and 506 F_2 families investigated for *YrSP*/*Yr5* and *YrSP*/*Yr7*, respectively). However, a more recent study showed that *YrSP* could actually be allelic to *Yr7* and *Yr5* (no susceptible progeny found among 208 and 256 tested F_3 families for *YrSP*/*Yr5* and

YrSP/*Yr7*, respectively[153]). Another study found only 2 and 4 susceptible F_2 among 522 and 506 families for *YrSP*/*Yr5* and *YrSP*/*Yr7* and the derived F_3 families were not tested[54], it is thus possible that these susceptible F_2 plants were escapes or phenotyping/genotyping mistakes. Therefore, it is still unclear whether *YrSP* is different from *Yr5* and *Yr7*. Elucidating the relationship between these genes is important for breeders, for example, to know whether all could be recombined in one variety. We thus addressed this question in this Chapter and Chapter 3.

2.1.1.3. Yr12

We mentioned in Chapter 1 that the separation between race-specificity of all-stage resistance genes and partial and broad-spectrum resistance conferred by APR genes was not clear. Several race-specific APR genes have been characterised for yellow rust disease in European cultivars[160]: *Yr11*, *Yr12*, *Yr13* and *Yr14*. The associated resistance response is a typical hypersensitive response starting to be expressed at tiller stage (Zadok's scale 20). It is thus slightly different in this aspect from seedling resistance which is expressed at an even earlier stage (1 leaf stage). However, the level of resistance is similar to that defined for seedling resistance. It is unknown whether this particular type of APR is encoded by NLRs, although the associated hypersensitive response is a hallmark of NLR-mediated resistance in plants.

Virulent *Pst* isolates were detected in the UK in the 1960s and 1970s for *Yr11*, *12*, *13* and *14*[160](https://ahdb.org.uk/ukcpvs). A later report stated that *Yr12* was still effective in China in 2002, although in combination with *Yr3a* and *Yr4a*[161]. On the other hand, virulence to *Yr12* was recorded in 2001-2004 in Ecuador[162]. There is very little information about these race-specific APR genes and they are not part of the World/European differential for yellow rust disease testing[163]. However, evidence of

Yr12-mediated resistance has been observed in the UK in the cultivar Armada (Simon Berry, personal communication). Armada was released in 1978 (Table 2-2) and carries *Yr3a* and *Yr4a* in addition to *Yr12*. Given that most of the tested current *Pst* isolates are virulent to *Yr3a* and *Yr4a* in the UK (https://ahdb.org.uk/ukcpvs), we hypothesized that the resistance observed in the field in Armada is mediated by *Yr12*.

Yr12 varieties are fairly old, from a wheat cultivar point of view. There is no report of wide deployment, at least for the UK varieties, from the 1990s to 2018 (NIAB-TAG Seedstats journal, NIAB-TAG Network). We could assume that because *Yr12* was never widely deployed in commercial cultivars, the selection pressure applied to *Pst* isolates virulent to this gene was never high enough to be selected for. Hence it is not surprising to currently observe *Yr12*-mediated resistance in the field given that it has yet to be widely deployed, a situation analogous to the fact that virulence to *Yr5* was observed once in 1984 but never since then (Table 2-1).

Table 2-2. Known *Yr12* varieties reported in the Genetic Resources Information System for Wheat and Triticale database (http://wheatpedigree.net/)

Name	Accession number	Locality	Year
CARSTENS-V	K-26401; PI-351206; PI-191311; CI-11768	Germany	1921
CARSTENS-VI	K-41875; AFRC-331; PI-180578; PI-282909	Germany	1940
NORD-DESPREZ	K-41873; PI-167419,351216; NGB-7037	France	1945
CARIBO	K-49832; CI-15177; AFRC-776,6628; AUS-12430	Germany	1968
MARIS-BEACON	K-51911; AFRC-572; PI-518814; AUS-12024	United Kingdom	1968
CYRANO	K-51609; AFRC-2832	Germany	1970
PAHA	K-49860; CI-14485; AFRC-2189	USA: Washington	1970
ANOUK	K-54100; K-53495	Belgium	1972
MEGA	K-54080; PI-410870; AFRC-927; AUS-19806	United Kingdom	1972
TAM-102	K-50447; CI-15283; AUS-15056; AFRC-853	USA: Texas	1973
PRIDE	AFRC-1580	United Kingdom	1974
FLEURUS	K-55352; PI-428518; PI-659576; ERGE-2345	France	1976
ARMADA	K-55338; PI-422222; PI-447041; AFRC-2573	United Kingdom	1978
OKAPI	K-56919; AFRC-6986; AUS-21148	Germany	1978
WAGGONER	K-55347; PI-447049; AFRC-3623; AFRC-2212	United Kingdom	1980
FRONTIER	K-56764; AFRC-3590	United Kingdom	1981

Cloning *Yr12* and developing gene-specific markers to assist its deployment in breeding programs alongside other race-specific genes would not only be highly valuable for breeders, but also from an academic point of view as it would answer the question whether race-specific APR genes could encode NLR immune receptors. Our working hypothesis in this thesis is that *Yr12* indeed encodes an NLR protein and we will further develop in this Chapter our progress towards achieving cloning of *Yr12*.

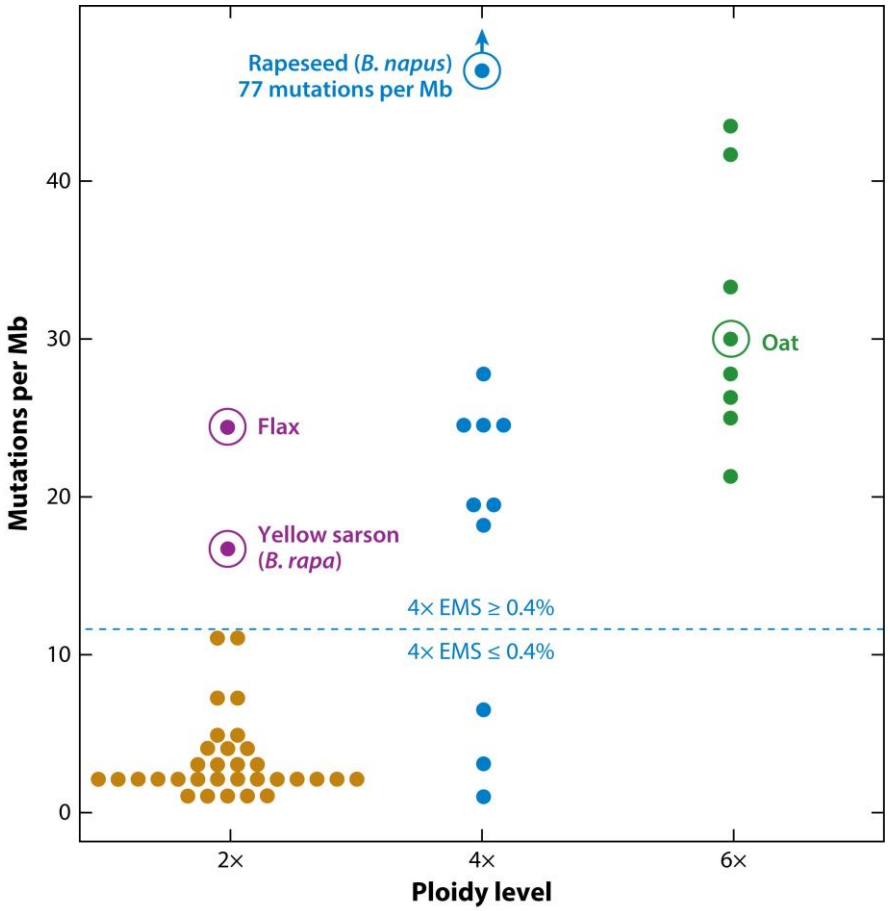
2.1.2. Forward genetic screens to identify yellow rust resistance gene loss of function mutants in wheat

In Chapter 1, we defined the concepts of forward and reverse genetics and provided examples of approaches that can be applied in wheat. Here, we will provide more details on forward genetic screens relying on chemically mutagenized populations in a well-characterised genetic background, as it is the approach we used to identify loss of function mutants in *Yr7*, *Yr5*, *YrSP* and *Yr12*. Once the cultivar of interest carrying the targeted gene is chosen, mutants are subsequently screened for the phenotype of interest (gain and/or loss of function). The way of identifying the responsible gene(s) evolved over the years thanks to various technological and methodological improvements[164] (discussed in Chapter 1). Usually, the phenotype is confirmed by genotyping mutant and wild-type individuals from a recombinant population and functional studies such as transgenic approaches can be pursued to further validate the candidate.

In the case of induced-mutations, there is a trade-off between achieving a high enough mutation density and the species' tolerance to such mutation rates. This is an important consideration for forward screens as the mutation density determines how many individuals are required to be screened to have a high enough probability of observing a

mutant in the gene of interest. If the mutation density in a population is low, a larger number of individuals is required to be investigated and vice-versa.

Polyploid species are more tolerant to high mutation rates, compared to diploids, because of the high functional redundancy between sub-genomes and are thus well suited for mutagenesis experiments as a smaller population is required to achieve saturation. The effect of polyploidy alone on mutation density was demonstrated in Arabidopsis[165]. The authors compared the effect of a given EMS treatment on a diploid and an autotetraploid derivative and showed that the autotetraploid could tolerate a higher mutation density with an improved survivability and fertility of the derived M_1 plants. For diploids, Uauy et al., (2017)[101] reported EMS-derived mutation densities ranging from 0.7 to 11.2 mutation(s)/Mb in *Cucumis melo* and *Arabidopsis thaliana*, respectively. In contrast, the highest mutation density reported in tetraploid species is 27.8 mutations/Mb in *Brassica napus* (notwithstanding a population with 77 mutations/Mb) whereas 41.7 mutations/Mb was observed in hexaploid wheat. Figure 2-2 illustrates the range of mutation densities reported so far in species with different ploidies.



Uauy C, et al. 2017.
Annu. Rev. Genet. 51:435–54

Figure 2-2. Mutation rates in mutant populations according to their ploidy level.

Diploid species are shown in gold, tetraploid species in blue and hexaploid in green. Flax and yellow sarson (purple), two diploid species that have undergone whole genome duplications within the last nine million years, have mutation rates comparable with those of tetraploid species and should be considered polyploids in an intermediate state of diploidization. This high spread of tetraploid values seems to originate in part by different dosages of the mutagen used in these species (see EMS threshold, blue dashed line).

Wheat is thus well-suited for EMS-based mutagenesis. However, the limitation of using polyploid species for forward genetic screens lies in the very same reason they can tolerate high mutation rates: functional redundancy between homoeologs. This limits the number of phenotypic mutants recovered from a forward screen because often mutants in a single homoeolog are indistinguishable from the wild type lines. Thus, single homoeolog mutants will not be "screened" and identified in the analysis. This is the worst-case scenario with complete functional redundancy among homoeologs, although often there is some subtle variation and dosage effect[166].

Most of the characterised resistance genes in wheat are dominant or semi-dominant[56]. Uauy et al., reported that among 135 rust resistance genes documented at the time of the study in the 2017 Catalogue of Gene Symbols for wheat, only 6.7% (9) are recessive whereas 26.3 % of reported resistance genes in barley are recessive. This can be linked to the functional redundancy described above, as it would require screening a M_2 line carrying a deleterious mutation in all homoeologs to identify such recessive genes in a forward genetic screen and the probability of such an event occurring is very low.

However, information from such forward genetic screens carried out in diploid relatives can be used in reverse genetics in wheat. For example, generation of exome-captured and sequenced TILLING populations in tetraploid[167] and hexaploid[107] wheat enabled selecting lines carrying mutations in the gene(s) of interest and its homoeologs to perform the necessary crosses and generate double and triple mutants before assessing the effect of the mutation on the phenotype[99,141]. Gene editing was also proven effective in generating triple mutants in the A, B and D genome homoeologs of *Mildew Locus O* (*MLO*[168], a rare case of recessive resistance gene from barley)[169]. Although these are

powerful approaches to conduct functional characterisation of genes in polyploid species, we will not pursue these in this thesis.

On the other hand, race-specific dominant resistance genes encoding NLR immune receptors rarely have true functional homoeologs. Thus, a homozygous mutation in the gene of interest is sufficient to see a loss of function phenotype in the population when screened with a normally avirulent pathogen isolate. Several resistance genes have been cloned in wheat using forward genetic screens in EMS-mutagenised populations: *Lr10*[170], *Tsn1*[171], *Sr22*/*Sr45*[137], *Sr33*[172]. Such approaches are thus suitable to target gene showing similar characteristics.

Yr7, *Yr5*, *YrSP* and *Yr12* are race-specific, dominant and drive a hypersensitive response in presence of the pathogen. These are the hallmarks of NLR-mediated resistance. Our working hypothesis is thus that all four encode NLR immune receptors. Given their dominant gene action and the absence of functional homoeologs for known NLRs, we hypothesized that it would be possible to detect loss of function mutants in an EMS-mutagenized population developed in a cultivar carrying the corresponding gene.

2.1.3. Summary and Disclaimer

In this Chapter we will describe how we obtained loss of function mutants for *Yr7* by screening an available EMS-mutagenized population in Cadenza and *Yr12* by generating a similar population in Armada. We will also briefly show the confirmation of already published *Yr5* loss of function mutants and describe additional plant materials used to clone *Yr7*, *Yr5* and *YrSP*. This provided us with the starting materials required for mutational genomics coupled with resistance gene enrichment sequencing (MutRenSeq, Chapter 3). Experiments for *Yr7*, *Yr5*, *YrSP* are published in Marchal et al., 2018[153].

Several experiments were carried out as part of a collaboration with Robert McIntosh (RM) and Peng Zhang (PZ) (University of Sydney), Paul Fenwick (PF) and Simon Berry (SB) (Group Limagrain UK). Table 2-3 shows the specific experiments carried out by our collaborators and we will refer to it thorough the Chapter. CU refers to Cristobal Uauy and CM to myself.

Table 2-3. Contribution of our collaborators to the work presented in this Chapter

Gene	Experiment	Contributed by
Yr12	Generating an EMS-mutagenised population in the cultivar 'Armada'	CM
	Arranging Year 1, 2 and 3 field trials for *Yr12*	SB, PF
	Genotyping M_3 lines for *Yr12* mutants	SB, PF
	Phenotyping *Yr12* mutants in the field (Year 1, 2 and 3)	PF, CM
	Selecting lines for MutRenSeq	CM, SB, CU
Yr7, *Yr5*, *YrSP* (AvocetS-Yr lines)	Identification of loss of function mutants for *Yr7*, *Yr5* and *YrSP* in the corresponding AvocetS-Yr7, AvocetS-Yr5, AvocetS-YrSP EMS-mutagenised populations	RM, PZ
Yr7	Screening the Cadenza TILLING population to identify Yr7 loss of function mutants	PF
	Progeny testing and F_2 populations generation to confirm loss of function mutants in *Yr7*	CM
	Selecting lines for MutRenSeq	CM, SB, CU
Yr5	Confirming *Yr5* loss of function phenotype in already published EMS-mutagenised line in Lemhi-Yr5 background	CM

2.2. Materials and methods

2.2.1. Plant materials and *Pst* isolates

We screened an available EMS-mutagenized population of the UK hexaploid cultivar 'Cadenza' to identify *Yr7* loss of function mutants[107]. The population is available through the John Innes Centre Germplasm Resource Unit: https://www.seedstor.ac.uk/search-browseaccessions.php?idCollection=24. We inoculated M_3 plants with *Pst* isolate 08/21 which is virulent to *Yr1, Yr2, Yr3, Yr4, Yr6, Yr9, Yr17, Yr32, YrRob, and YrSol*[173]. We

used the following nomenclature for the Cadenza lines, with Cad127 standing for Cadenza0127. We screened two independent batches of 500 lines (1,000 Cadenza mutant lines in total; sections 2.3.1.1 and 2.3.1.3). The two screens were performed in the exact same conditions to ensure the consistency of loss of function mutants identified in each.

To clone *Yr5*, we used published EMS-mutants in the 'Lemhi-Yr5' background[174]. To identify *Yr12* loss of function mutants, we carried out EMS-mutagenesis in UK cultivar 'Armada' (provided by Group Limagrain UK).

EMS-derived mutants from Avocet-Yr7, Avocet-Yr5 and Avocet-YrSP were developed and screened with adequate *Pst* isolates by Peng Zhang and Robert McIntosh (University of Sydney). *Pst* pathotypes 108 E141A+ (University of Sydney Plant Breeding Institute Culture no. 420), 150 E16A+ (Culture no. 598) and 134 E16A+ (Culture no. 572) were used to evaluate *Yr7*, *Yr5*, and *YrSP* mutants, respectively. Three *Yr7* loss of function lines, four *Yr5* lines and four *YrSP* lines were identified in Avocet-Yr7, Avocet-Yr5 and Avocet-YrSP populations, respectively. All lines are described in Appendix 8-3.

2.2.2. EMS mutagenesis in the cultivar Armada and mutant selection

We tested four ethyl methanesulfonate (EMS, 62-50-0 Sigma Aldrich) concentrations: 0.6, 0.7, 0.8 and 0.9 % (v/v) and sampled ~ 3,000 M_0 seed per concentration. We incubated M_0 seeds in a 10 % Tween20 solution for 15 mins on a roller bar shaker. Then we washed the seeds with water four times five minutes to eliminate Tween20. Following this we added the EMS solution and incubated the seeds on the roller bar shaker for 18 hours with gentle shaking to avoid seed to break. Finally, we rinsed the seeds five times 15 mins with water before placing them under running water for one hour to eliminate

as much EMS as possible. We secured the bottle with a cheesecloth to avoid the seeds washing out at this step.

Figure 2-3 summarises our approach to select *Yr12* loss of function mutants in the field. We pre-germinated M_1 seed on water-imbibed filter paper in trays placed at 4˚C and transferred them to soil when they were producing three roots. Trays were kept in a glasshouse with no additional lighting or heating. Once M_1 seedlings reached two to three leaves stage we transferred them in a controlled-environment room set to 8˚C for eight weeks to allow for vernalisation, followed by a week at 12.5˚C for acclimation before being finally grown in a glasshouse with sunlight control but without heating control for seed production.

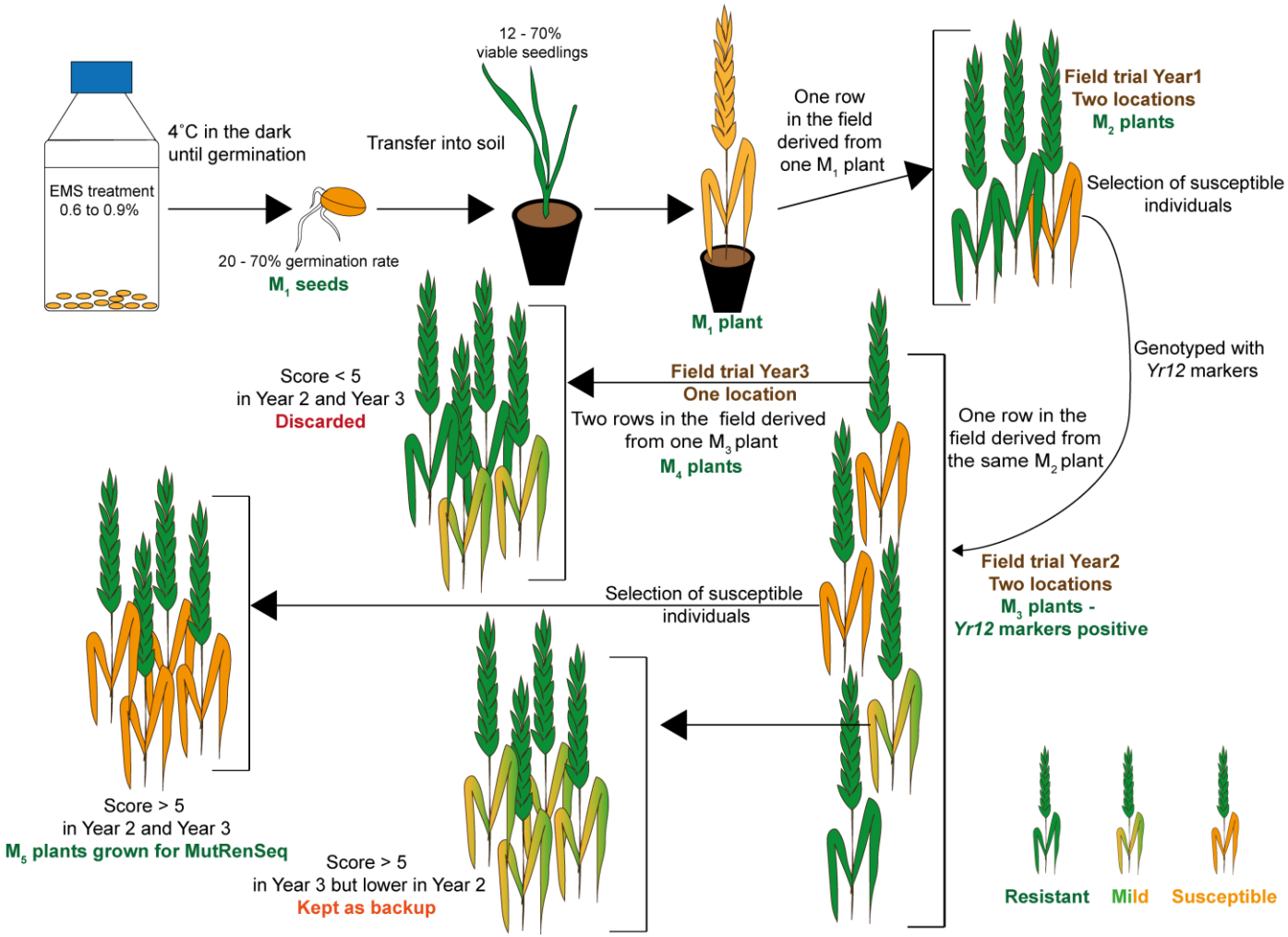


Figure 2-3. Schematics showing the workflow to identify *Yr12* loss of function mutants for MutRenSeq analysis.

Since *Yr12* confers adult-plant resistance in the field, we thus tested the mutant population in the field during summer 2017 (Year 1). We drilled one row per original M_1 plant with the wild-type Armada as a positive control to monitor for *Yr12*-mediated resistance in the field at the time of the screen at two locations (Rothwell, 53°28'50.7"N 0°15'11.4"W and Osgodby 53°25'11.5"N 0°23'03.7"W). Because our working hypothesis was that *Yr12* encodes an NLR immune receptor and that these genes are usually dominant, a homozygous mutation would be needed to knock-out the gene and display a susceptible phenotype. We thus expected that the susceptible phenotype would segregate in the M_2 progeny.

For this first screen we scored presence of yellow stripes and pustules as susceptible and their absence as resistant. This allowed us to be inclusive and identified all potential *Yr12* loss of function mutants. We tagged susceptible M_2 plants and threshed all spikes independently. These were subsequently genotyped to ensure the presence of the Armada haplotype across the wider *Yr12* region (markers are shown in Appendix 8-3). Each selected M_3 was multiplied in the glasshouse to produce enough material for the following field trial (2018, Year 2).

We conducted the 2018 (Year2) trial at two locations (Rothwell, 53°28'50.7"N 0°15'11.4"W and Osgodby 53°25'11.5"N 0°23'03.7"W). For this round of selection in the field, we harvested all derived M_4 seeds but prioritised the ones with i) the highest infection type (IT, Table 2-4), ii) no apparent segregation of the susceptible phenotype in the progeny, and iii) consistent traits at the two locations tested.

Table 2-4. Infection type (IT) scores for yellow rust disease in field plots.

Score	% infection	Comments
0	0	No evidence of yellow rust disease
1	0.1	1 stripe per tiller
2	1	2 stripes per tiller
3	5	most tillers infected but some top leaves uninfected
4	10	All leaves infected but leaves appear green overall
5	15	4+
6	25	Leaves appear 1/4 infected and 3/4 green
7	35	Leaves appear 1/3 infected and 2/3 green
8	50	Leaves appear 1/2 infected and 1/2 green
9	75	Very little green tissue visible
10	100	Leaves are dead

In parallel, we tested all derived M_4 plants in the field in one location (Rothwell) the following year (2019, Year 3) and sowed a subset in the glasshouse, planning for the end of vernalisation coinciding with when plants in the field would be ready for scoring. For a given line, M_4 plants used for genotyping and M_4 plants sown in the field the same year are derived from the same M_3 individual. Thus, providing the phenotype is no longer segregating in the field plots, we can save time with selecting susceptible M_4 in the field and directly sampling leaf tissue from young glasshouse-grown plants derived from the same M_3 individual. Progenies showing an IT > 5 in both Year 2 and Year 3 were selected for MutRenSeq. Progenies showing an IT > 5 in Year 3 and lower in Year 2 were kept as backup. Progenies showing IT < 5 in both Year 2 and Year 3 were discarded.

I did the EMS mutagenesis experiment and monitored the seedlings until they reached 2-3 leaves stage. After that the population was maintained in Rothwell and Woolpit stations (Limagrain). All field trials were designed by Cristobal Uauy, Simon Berry, Paul Fenwick and myself and carried out by Simon Berry and Paul Fenwick in Rothwell station (Limagrain). Simon Berry, Paul Fenwick and myself did the yellow rust disease scoring in the field for all trials and selected the lines to be advanced to the next

generation and ultimately used for the MutRenSeq experiment. Genotyping of the M_3 plants was performed in Limagrain Genotyping Lab in Clermont Ferrand (France).

2.2.3. Seedling tests to identify *Yr7* loss of function mutants

Yr7 resistance is present at all stages of plant growth and can thus be screened for at seedling stage. We screened M_3 plants from the EMS-mutagenized population in Cadenza for *Yr7* loss of function mutants. Cadenza is a cultivar released in 1992 that was a prevalent parental line in UK breeding programs. It is known to carry at least one additional yellow rust resistance gene, *Yr6*, in addition to *Yr7*. Hence we chose PST 08/21 (*Yr1, Yr2, Yr3, Yr4, Yr6, Yr9, Yr17, Yr32, YrRob, and YrSol*[173]) to be able to discriminate between the two resistance genes as this isolate is virulent to *Yr6*, but avirulent to *Yr7*.

Paul Fenwick tested 1,000 lines in total and sowed four seeds per line in two independent batches of 500 lines each. Plants were grown in 192-well trays in a confined glasshouse with no supplementary lights or heat. Inoculations were performed at the one leaf stage (Zadoks 11) with a talc-urediniospore mixture. Trays were kept in darkness at 10 °C and 100 % humidity for 24 hours. Infection types (IT) were recorded 21 days post-inoculation (dpi) following the Grassner and Straib scale[175]. Identified susceptible lines were progeny tested (twelve to 16 plants per line) in similar conditions as described above to confirm the reliability of the phenotype.

2.2.4. Seedling tests to confirm published *Yr5* mutants

Information about generating the mutant population in Lemhi-Yr5 background and *Yr5* loss of function mutant selection was described in McGrann et al., 2014[174]. These mutants were selected with the *Pst* isolate PST 81/20 that is virulent to Lemhi, AvocetS

and Chinese Spring but avirulent to Lemhi-Yr5 and spelt wheat cultivar Album (*Yr5* donor). The authors classified the lines in two groups depending on their segregation pattern in F_2 progenies derived from a cross between a mutant line and AvocetS. AvocetS does not carry *Yr5* and is susceptible to PST 81/20, thus it is expected in that both F_{1s} and F_{2s} are all susceptible to the tested *Pst* isolate. However, the authors observed segregation of the susceptible phenotype in several F_2 families, indicating that the mutation leading to the susceptible phenotype in Lemhi-Yr5 could be complemented by AvocetS. Thus, this mutation is not in *Yr5*, as we stated above that AvocetS does not carry *Yr5*. Table 2-5 summarises the Lemhi-Yr5 mutant lines generated in this work and the corresponding hypothesis regarding the causal mutation.

Table 2-5. Description of Lemhi-Yr5 mutant lines from McGrann et al., 2014.
IT scores correspond to the Grassner and Straib scale with 0 indicated full resistance and 4 full susceptibility. "n" stands for necrotic spots.
Hypothesis regarding the nature of the causal mutation derives from the segregation ratio of resistant:susceptible plants in the F_2 progenies: if 100% of the progenies are susceptible, then we assumed the causal mutation is in *Yr5* because AvocetS background could not complement *Yr5* loss in the mutant line. However, if some resistant lines were present in the progeny, then AvocetS could complement the loss of resistance in the mutant line. Thus the causal mutation cannot be located in *Yr5*. We selected lines for MutRenSeq based on this hypothesis (Y for Yes and N for NO).
Lem18 stands for Lemhi-Yr5 mutant line number 18, this nomenclature will be used thorough the thesis

Line	**Generation**	**F_1 IT (0 to 4)**	**Phenotype segregation in F_2 progenies (resistant: susceptible)**	**Causal mutation hypothesis**	**Sent for RenSeq in the present thesis (see Chapter 3)**
Lem18	M_3	4	20:214	Locus independent from *Yr5* but important for resistance	N
Lem90	M_3	4	21:189	Locus independent from *Yr5* but important for resistance	N
Lem94	M_3	0^{nn}	132:84	Locus independent from *Yr5*-mediated resistance	N
Lem95	M_3	4	0:234	Yr5	Y
Lem98	M_3	4	100:122	Locus independent from *Yr5* but important for resistance	N
Lem99	M_3	4	12:218	Locus independent from *Yr5* but important for resistance	N
Lem115	M_3	4	0:228	Yr5	Y
Lem241	M_3	4	0:212	Yr5	Y
Lem287	M_3	4	0:216	Yr5	Y
Lem387	M_3	4	0:220	Yr5	Y
Lem474	M_3	4	0:218	Yr5	Y
Lem500	M_3	4	0:228	Yr5	Y
Lemhi-Yr5	-	$0^n/1^n$	170:64	-	Y
Lemhi	-	-	-	-	Y

We obtained the M_3 seeds from this work and tested them against PST 81/20, which is avirulent to Lemhi-Yr5 but virulent to Lemhi, to confirm the phenotype. We performed the seedling tests following the same protocol as for *Yr7* loss of function mutants, although here eight plants per line were tested. We selected mutants with the highest infection type score based on both the publication and the confirmation test (seven lines).

2.3. Results

2.3.1. Investigating available materials to identify loss of function mutants for *Yr7*, *Yr5* and *YrSP*

The following experiments are summarised in Marchal et al., 2018[176]. Here we provide detailed information on each of the plant materials that have been used to identify loss of function mutants in the targeted genes. We summarised our approach to validate the *Yr7*-loss of function mutants in the TILLING population in Cadenza in Figure 2-4. This figure refers to the results presented in sections 2.3.1.1 and 2.3.1.3.

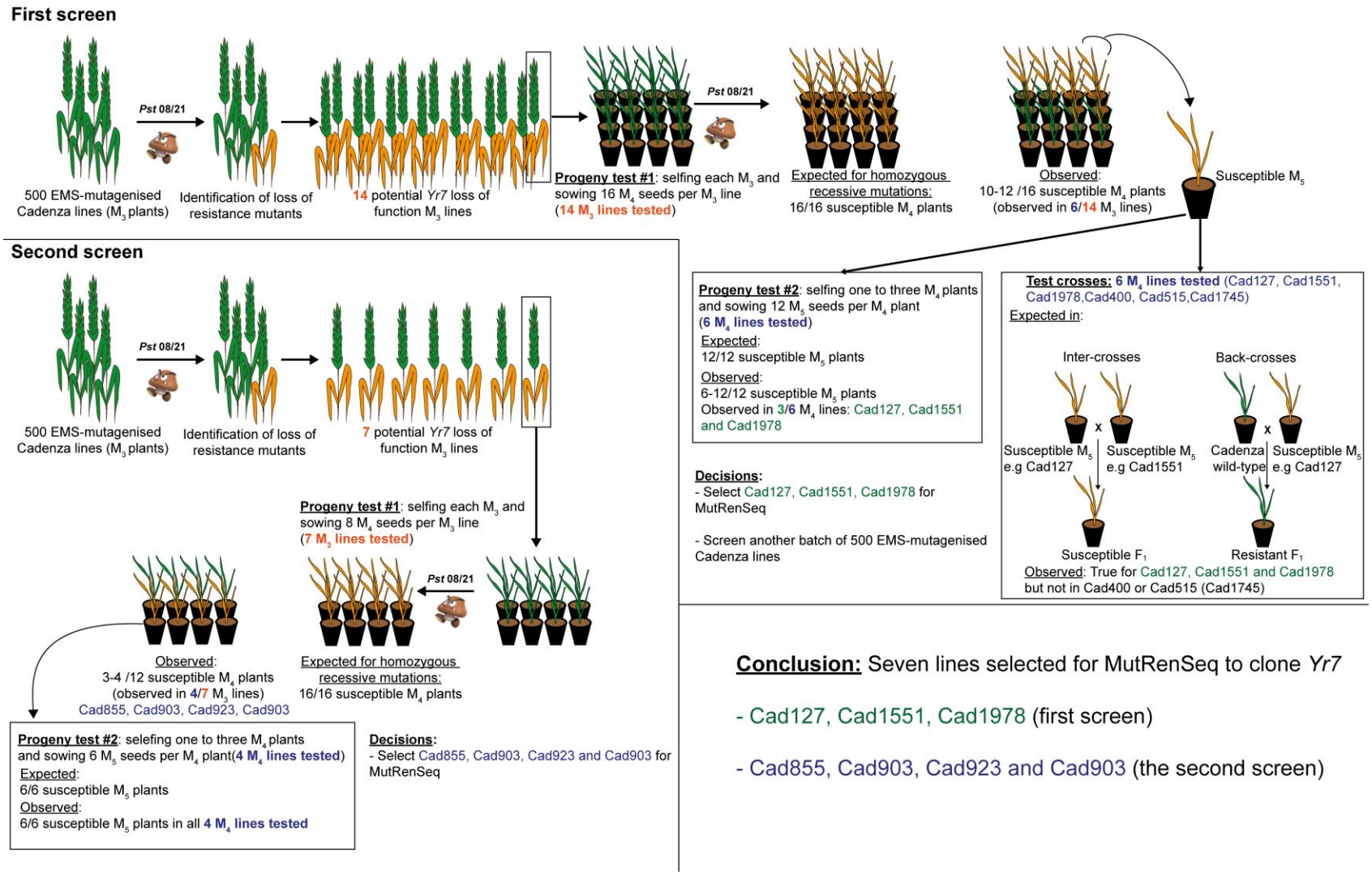


Figure 2-4. Summary of the different steps we followed to confirm the phenotype of the Yr7-loss of function mutants identified in the Cadenza EMS mutagenised population

2.3.1.1. Screening an available TILLING population in Cadenza allowed identification of loss of function mutants in Yr7 – first screen

We screened a first batch of 500 M_2 lines from the EMS-mutagenized population in Cadenza[107] and identified 14 susceptible independent lines. Each Cadenza mutant line was progeny tested to investigate the segregation of the susceptible phenotype in the derived plants (Table 2-6). Since *Yr7* is a dominant gene, we hypothesised that plants with a susceptible phenotype should carry homozygous recessive mutations in *Yr7*. We thus hypothesised that all $M_{(x+1)}$ progenies from the susceptible M_X plants should be susceptible. However, none of the tested lines confirmed this (Table 2-6). Nonetheless we observed a clear separation between lines showing 9 to 15 susceptible lines out of 16 plants tested and lines showing only a few (0 to 4) susceptible plants in their progeny. We thus decided to select the six lines having the majority of susceptible plants in their progeny (highlighted in green in Table 2-6), noting that resistant plants could be escapes. Cad665 was not retained due to the lower infection type observed when compared to the other susceptible lines.

We performed backcrosses to Cadenza wild-type for the six selected lines to determine the phenotype of the resulting F_1 plants. For all crosses we obtained resistant F_1 plants (Table 2-6). This agrees with our working hypothesis that the *Yr7* loss of function phenotypes are based on homozygous recessive mutations.

We next performed inter-crosses between susceptible lines to determine whether they belonged to the same complementation group. All crosses involving Cad400 and Cad515 produced resistant F_{1s}, whereas crosses involving Cad127, Cad1978 and Cad1551 produced susceptible F_{1s} when crossed together and wild-type F_{1s} when crosses to Cad400 and Cad515. Thus, we assigned Cad127, Cad1551 and Cad1978 to the same

complementation group and likely to carry mutations in the same gene, *Yr7*. Cad400 and Cad515 belonged to two different complementation groups and we hypothesised that they could carry mutations in different genes that are required for *Yr7*-mediated resistance.

We repeated the progeny test and confirmed the previous results. Although infection types were lower than in the first test, we still identified a good proportion of susceptible progenies for 2/3 of the tested M_5 plants from Cad127, one tested from Cad1551 and two tested from Cad1978 whereas all Cad400 and Cad515 progenies were resistant (Figure 2-5). We did not retain Cad1745 because all the progenies had strong developmental issues and most of them did not germinate. This can be due to other background mutations that were selected with the susceptible phenotype.

Table 2-6. Summary of the phenotype confirmation of the Cadenza mutant lines
Three out of the 14 identified lines were selected based on the described experiments. First, each M_4 line was progeny tested to study the segregation pattern of the susceptibility phenotype in the descendance.
Here we defined any presence of pustule (1^- to 4) as susceptibility. Inter-crosses between susceptible lines and backcrosses to Cadenza wild-type informed on whether mutants belonged to the same complementation group. A second progeny test was carried out on the M_5 to validate results. Green highlight depicts validation of the line for the given test and agreement with the hypothesis susceptible phenotype caused by a recessive mutation.

Susceptible line	**#Susceptible plants (16 tested plants per M_4) (score range)**	**(hypothesis)/ Mutation status**	**F_1: inter crosses with other susceptible lines (3 F_1 plants)**	**F1: back-crosses to wild type (3 F_1 plants)**	**#Susceptible plants in second test from M_5 (12 plants tested per M_5) - (susceptible phenotype range)**
Cad188	4 (1^-)	HET	n.t	n.t	n.t
Cad127	10 (1-3)	(dom?) HET	- Cad127 x Cad1551: susceptible - Cad127 x Cad515: resistant - Cad127 x Cad400: resistant - Cad127 x Cad 1978: susceptible	Resistant phenotype	$Cad127_1$: 1 (1^-) $Cad127_2$: 6 (1^-) $Cad127_3$: 6 (1^-)
Cad248	4 (1^-)	HET	n.t	n.t	n.t
Cad339	0	mutation loss or escape	n.t	n.t	n.t
Cad392	1 (1^-)	mutation loss or escape	n.t	n.t	n.t
Cad400	15 (1^--3)	HOM	- Cad400 x Cad515: resistant - Cad400 x Cad1551: resistant - Cad127 x Cad400: resistant	Resistant phenotype	0
Cad421	1	mutation loss or escape	n.t	n.t	n.t
Cad515	10 (1^--2)	(dom?) HET	- Cad515 x Cad400: resistant - Cad515 x Cad 1978: resistant	Resistant phenotype	$Cad515_1$:0/12

Susceptible line	#Susceptible plants (16 tested plants per M_4) (score range)	(hypothesis)/ Mutation status	F_1: inter crosses with other susceptible lines (3 F_1 plants)	F1: back-crosses to wild type (3 F_1 plants)	#Susceptible plants in second test from M_5 (12 plants tested per M_5) - (susceptible phenotype range)
			- Cad515 x Cad 127: resistant		
Cad665	9 (1^--1)	HET	n.t	n.t	n.t
Cad667	2 (1^-)	mutation loss or escape	n.t	n.t	n.t
Cad1551	10 (1^--4)	(dom?) HET	- Cad127 x Cad1551: susceptible - Cad400 x Cad1551: resistant	Resistant phenotype	$Cad1551_1$:9 (1-2)
Cad1745	12 (1^--2)	(dom?) HET	n.t (did not germinate)	n.t	Very small plants
Cad1746	2 (1^-)	mutation loss or escape	n.t	n.t	n.t
Cad1978	7/13 (1^--4)	(dom?) HET	- Cad515 x Cad 1978: resistant - Cad1551 x Cad 1978: susceptible		$Cad1978_1$:5/9* (1-3) $Cad1978_2$:11 ($1\text{-}2^+$)

*Three individuals did not germinate. "n.t" stands for not tested

Given the consistent results in progeny tests, the recessive nature in F_1 hybrids and complementation results, we focused on Cad127, Cad1551 and Cad1978 for MutRenSeq. We extracted DNA from leaves of one susceptible plant per selected family. This plant was then crossed to Cadenza wild-type and we advanced the derived F_1 to F_2 generation. These F_2 plants will be used to confirm genetic linkage between candidate mutations and the *Yr7* loss of function phenotype.

2.3.1.2. *Calculating the probability of lines sharing a mutation by chance*

We selected three lines for MutRenSeq, but is that enough to identify relevant candidate gene(s) in a MutRenSeq experiment? We explored this based on previous work using the following formula to determine how likely it is for a certain number of mutants to have mutations in the same gene by chance[134]:

$$P_m = l * M$$

$$P_w = P_m^x$$

P_m: Mutation probability; l: contig length; M: mutation density; x: number of independent lines; P_w: Probability of mutated contig

We used N_{50} calculated from *de novo* assemblies from RenSeq data to simulate contig length (l). N_{50} from such data varied from 1,745 to 2,864 bp in a previous study[177]. EMS-type mutation (G to A and C to T) density (M) per Cadenza line was estimated as 33 mutations per Mb[107]. Thus, assuming GC content in NLR is similar to all exons in general, we estimated that the probability for three mutant lines having a mutation in the same RenSeq contig by chance would range between 1.7E-4 to 8.44E-3. Given that the number of contigs associated to NLRs ranged between 7,117 and 16,905 in *de novo* assemblies generated from RenSeq data[177], this means that two to 142 contigs could carry mutations in the same gene by chance if three mutant lines are investigated. Therefore, we decided to screen another batch of 500 lines to identify more loss of function mutants to strengthen the power of the analysis.

2.3.1.3. Screening an available TILLING population in Cadenza allowed identification of loss of function mutants in Yr7 – second screen

We identified seven additional putative susceptible M_3 lines in the second screen (Methods section 2.2.3, Table 2-7). As before, we observed segregation of the susceptible phenotype in the M_3 plants derived from a single M_2 plant. However, we lost the information regarding which plant out of the four tested was susceptible. We therefore progeny tested all four tested M_3 plants that were sown for the second screen. This explains why most of the lines had very few or not susceptible plants in the progeny test (Table 2-7).

As we observed earlier for the three previously selected lines, we did not reported susceptibility in all progenies for a given line. We thus selected families for which at least 3/8 tested progenies were susceptible: Cad855, Cad903, Cad923, Cad1034, Cad1105. The phenotype was confirmed in an additional pathology test prior sending DNA sample for RenSeq (Figure 2-5).

Table 2-7. Progeny test of all four M_3 lines from second screening of Cadenza EMS mutants with Pst isolate 08/21.
Because we lost the information regarding which of the four tested plants were susceptible, we had to progeny test all of them. Progeny tests were conducted the same way as on Table 2-6.

Susceptible line (#susceptible plants per 8 seedlings)	**#Susceptible in the progeny (8 plants tested per M_4)**	**Hypothesised mutation status for M_4 parent**
Cad855 (4)	3	HET
	1	Loss of mutation or escape from initial screen
	1	Loss of mutation or escape from initial screen
	1	Loss of mutation or escape from initial screen
Cad903 (4)	3	HET
	2	HET
	2	HET
	1	Loss of mutation or escape from initial screen
Cad923 (4)	4	HET
	3	HET
	3	HET
	2	HET
Cad1034 (3)	3	HET
	1	Loss of mutation or escape from initial screen
	1	Loss of mutation or escape from initial screen
Cad1105 (3)	2	HET
	0	Loss of mutation or escape from initial screen
	0	Loss of mutation or escape from initial screen
Cad1154 (2)	1	Loss of mutation or escape from initial screen
	1	Loss of mutation or escape from initial screen
Cad1216 (3)	0	Loss of mutation or escape from initial screen
	0	Loss of mutation or escape from initial screen
	0	Loss of mutation or escape from initial screen

Using the formula outlined in Section 2.3.1.2, we calculated the probability of having a gene mutated in seven independent lines by chance in a RenSeq dataset. We found that the probability would range between 2.2E-9 and 6.7E-8. This means that between 3.7E-5 and 1E-3 contigs would carry mutations in the same contig by chance if seven independent mutant lines are used. We thus decided that seven lines is likely to be enough to identify relevant candidates contigs

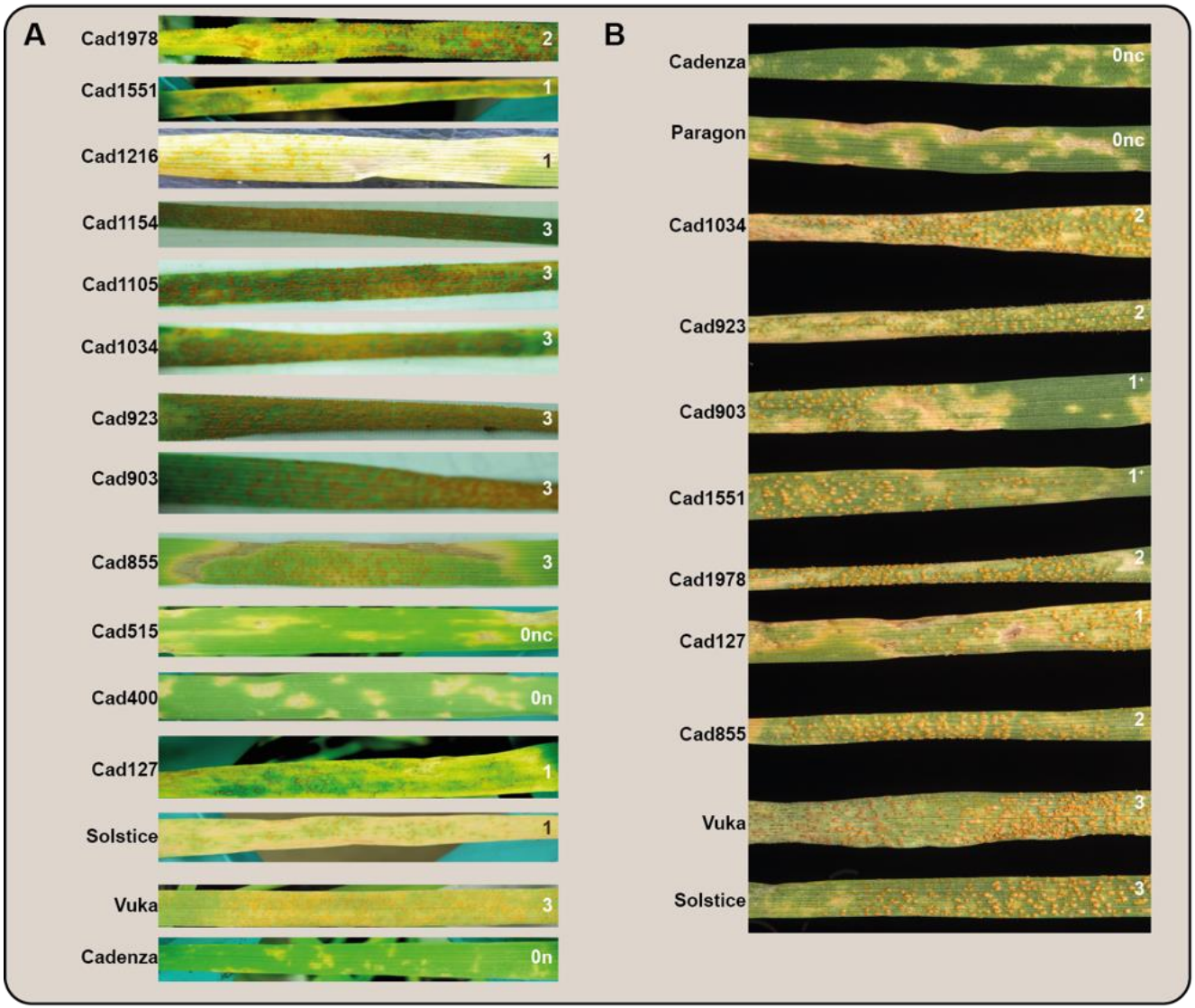


Figure 2-5. Yellow rust disease scoring of Cadenza mutant lines.
Line identifier is given on the left of the leaf picture and IT score on the right. We used Grassner and Straib scale for scoring. We called susceptible a line where *Pst* was able to complete its life-cycle (presence of pustules). Vuka and Solstice were used as positive controls for inoculation (both susceptible to PST 08/21 and do not carry *Yr7*) and Cadenza as negative control (resistant to PST 08/21 and carries *Yr7*).
A. Yellow rust disease scoring during the first progeny test on susceptible Cadenza mutant lines identified in the initial screen. Cad400 and Cad515 lost the susceptible phenotype during this first test and Cad1216 and Cad1154 lost the susceptible phenotype during the second progeny test. Cad1105 had very few susceptible progenies in the second test. These lines were thus not sent for RenSeq.
B. Final seedling test with PST 08/21 on Cadenza mutants and wild-type. Paragon is a Cadenza-derivative and carries *Yr7*

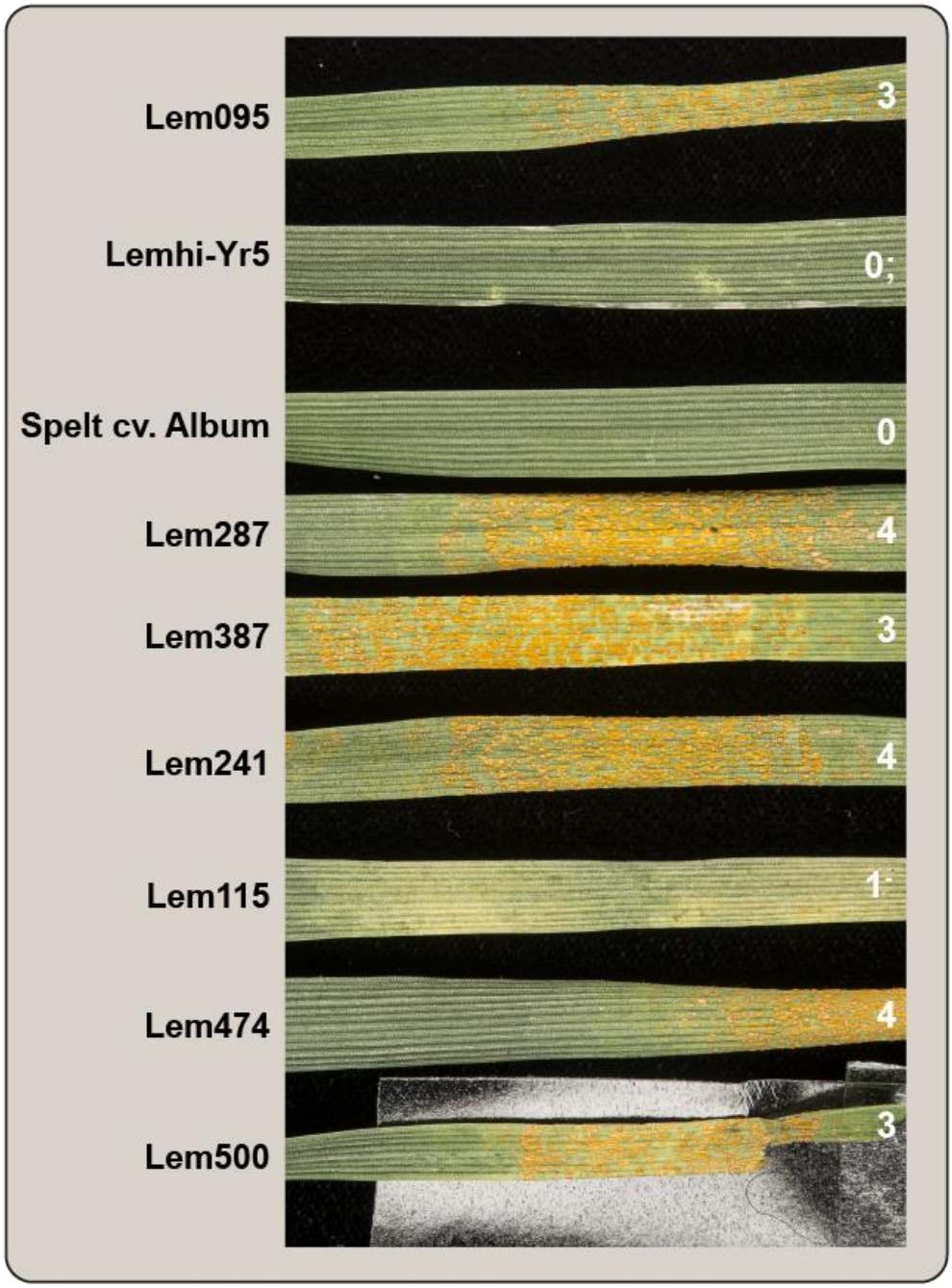
2.3.1.4. *Confirming phenotype of published loss of function mutants in Yr5*

McGrann et al. (2014)[174] published *Yr5* loss of function mutants in Lehmi-Yr5. From their work, we selected lines with the highest infection type and performed the seedling tests on seven lines as described in the Methods section 2.2.4. We obtained comparable scores as in the publication apart from Lem115, which was lower than 4 in our test (Table 2-8). We still selected this line for RenSeq based on the strong genetic evidence provided in the published work and made a note of its lower infection type when analysing the data. In total, we selected seven independent lines, which we previously argued would be enough to identify relevant candidates for the gene of interest.

Table 2-8. Comparison of seedling tests from published Lemhi-Yr5 mutants with our seedling tests on the same lines

Line	Score in McGrann et al., 2014	Score in our seedling test
Lem95	4	3
Lem115	4	1^-
Lem241	4	4
Lem287	4	4
Lem387	4	3
Lem474	4	4
Lem500	4	3
Lemhi-Yr5	n	0;
***Triticum spelta* cv. Album**	0;	0

Figure 2-6. Yellow rust disease scoring of M_3 susceptible Lemhi-Yr5 mutant originally identified in McGrann et al., 2014.
Line identifier is given on the left of the leaf picture and score on the right. We used Grassner and Straib scale for scoring. Lemhi-Yr5 and spelt cultivar Album were used as negative control for inoculation (both carry *Yr5* and are resistant to PST 81/20



2.3.2. Developing an EMS-mutagenized population in the cultivar Armada to identify *Yr12* loss of function mutants

2.3.2.1. Development of an EMS-mutagenised population in Armada

Armada carries the adult plant resistance gene *Yr12*. Because of the race-specific nature of *Yr12*-mediated resistance, we hypothesized that *Yr12* encodes an NLR immune receptor. We thus developed an EMS-mutagenized population in the cultivar Armada and carried out a MutRenSeq approach to clone *Yr12*. We tested the population through three consecutive field trials (one per plant generation) to identify *Yr12* loss of function mutants (see 2.2.2).

Wheat's polyploidy allows for a higher mutation density than for diploid plants due to the functional redundancy of the homoeologs[178]. Achieving a high mutation density is important for forward genetic screens to ensure that all genes have a chance to carry a mutation in the population. Based on the EMS-concentrations used to develop the Cadenza population[179], we tested four different EMS concentrations and assessed corresponding germination and seedling viability (Table 2-9). As expected, germination rate and seedlings viability decreased with increasing concentrations of EMS. More than 90 % of the viable seedlings produced seed, thus sterility was not a major issue if the plants survived past seedling stage. A total of 1,451 M_2 lines were available to test in the field the following year for *Yr12* loss of function. That is more than the total of Cadenza lines we screened (1,000) to identify seven confirmed *Yr7* loss of resistance mutants. Providing Armada and Cadenza respond in a similar way to EMS, we assumed that testing 1,451 lines would allow us to identify enough *Yr12* loss of function mutants to identify relevant candidate genes.

Table 2-9. Summary of the EMS mutagenesis experiment in Armada.

Concentration	#M0 seed	%M0 infected by fungi or broken	#Sown M0 seed	% germinated seed	# viable seedlings	% viable seedlings	# fertile	% fertile
0.60%	3000	about 50	1063	70.8	820	77.1	794	96.8
0.70%	3000	about 50	808	53.9	475	58.8	445	93.6
0.80%	3000	about 50	658	43.9	175	26.6	173	98.8
0.90%	3000	about 50	326	21.7	40	12.3	39	97.
Total	12000		**2855**	**23.8**	**1510**	**52.9**	**1451**	**96.1**

We sowed the 1,451 lines (one row per line) in two locations and scored the lines for presence/absence of yellow rust disease symptoms. We identified 59 susceptible lines in Rothwell and 36 in Osgodby with 18 lines susceptible at both sites. Because the disease was overall very low in Osgodby, we decided to discard all lines that were susceptible in Osgodby only. Out of the 59 remaining lines, we discarded another 6 because their overall plant phenotype differed from Armada and they were thus likely seed contamination. At the end of the growing season we harvested 39 out of the 56 lines to grow M_3 plants for genotyping. The 14 remaining lines showed a development delay and were not ready to be harvested at that time. We threshed between one to six individual spikes for each of the 39 plants and we genotyped each derived progeny independently. This led to a total of 96 independent progenies to evaluate (from 39 original mutant lines).

2.3.2.2. Identification of Yr12 loss of function mutants through three field trials

Appendix 8-3 shows the genotyping results for the 96 progenies derived from 39 identified susceptible lines in Year1. We used four markers linked to *Yr12* to select for lines having Armada haplotype in this region and an additional set of 44 markers spread

across all chromosomes to ensure we selected for Armada-like lines only. We tested nine M_3 plants on average per M_2 line and included Armada wild-type as a positive control and Vuka as negative control. Overall there was no genotypic difference between different spikes from the same line for the tested markers. Most of the tested lines were consistent with Armada-like profile (81/96) so we grew the corresponding plants to produce enough seed for Year3 field trial. We discarded 8/39 lines that did not show an Armada-like genotype or showed too many heterozygous plants for each tested marker (Appendix 8-3, "Discard" column).

In parallel the remaining bulks derived from the 39 susceptible lines identified in Year1 were sown for Year2 trial at the same two locations as in the first trial. Overall disease intensity was lower in Osgodby than in Rothwell. This time we used the key shown in Table 2-4 to select only for the most susceptible plants (score > 5). We identified 16 lines corresponding to this criterion in both sites. Note that this trial does not consider the 81 families derived from the 31 lines that passed the genotyping test. Indeed, because we did not have enough seed material to test these in Year2, we had to self each of the 81 lines and the seeds were sown in the Year3 trial.

In Year3 trial we sowed two row per independent spike genotyped (M_4 plants) only in Rothwell as it seemed that Osgodby was showing lower disease pressure than this site. The susceptible phenotype was not segregating any longer in most of the lines: only 3/81 of genotyped lines were scored as HETs (Appendix 8-3). There was little variation between different spikes coming for the same M_2 plant. However, in a few cases we observed important variation. For example, the ARM003 spikes scores ranged from 0 (fully resistant for ARM003A) to 7 (susceptible for ARM003B) and the same was observed for ARM023 and ARM033. We also noted some variation within the different

progenies derived from one given spike for several lines: ARM041C scores ranged from 2 (ARM041C-004) to 5 (ARM041C-003). Thus, even if the genotyping results were consistent across spikes derived from the same plant, the phenotypes were very different for a small number of lines. This is important to note as EMS mutation profiles are not necessarily identical in all tillers of the same M_2 plant.

From the original 96 lines, we selected 12 lines for MutRenSeq based on the consistency of their phenotype between Year1, Year2 and Year3 (Priority: IT score > 5 in Rothwell in Year2 and Year3; scored as susceptible in both sites in Year1, "x" on Appendix 8-3). Figure 2-7 shows the selected lines.

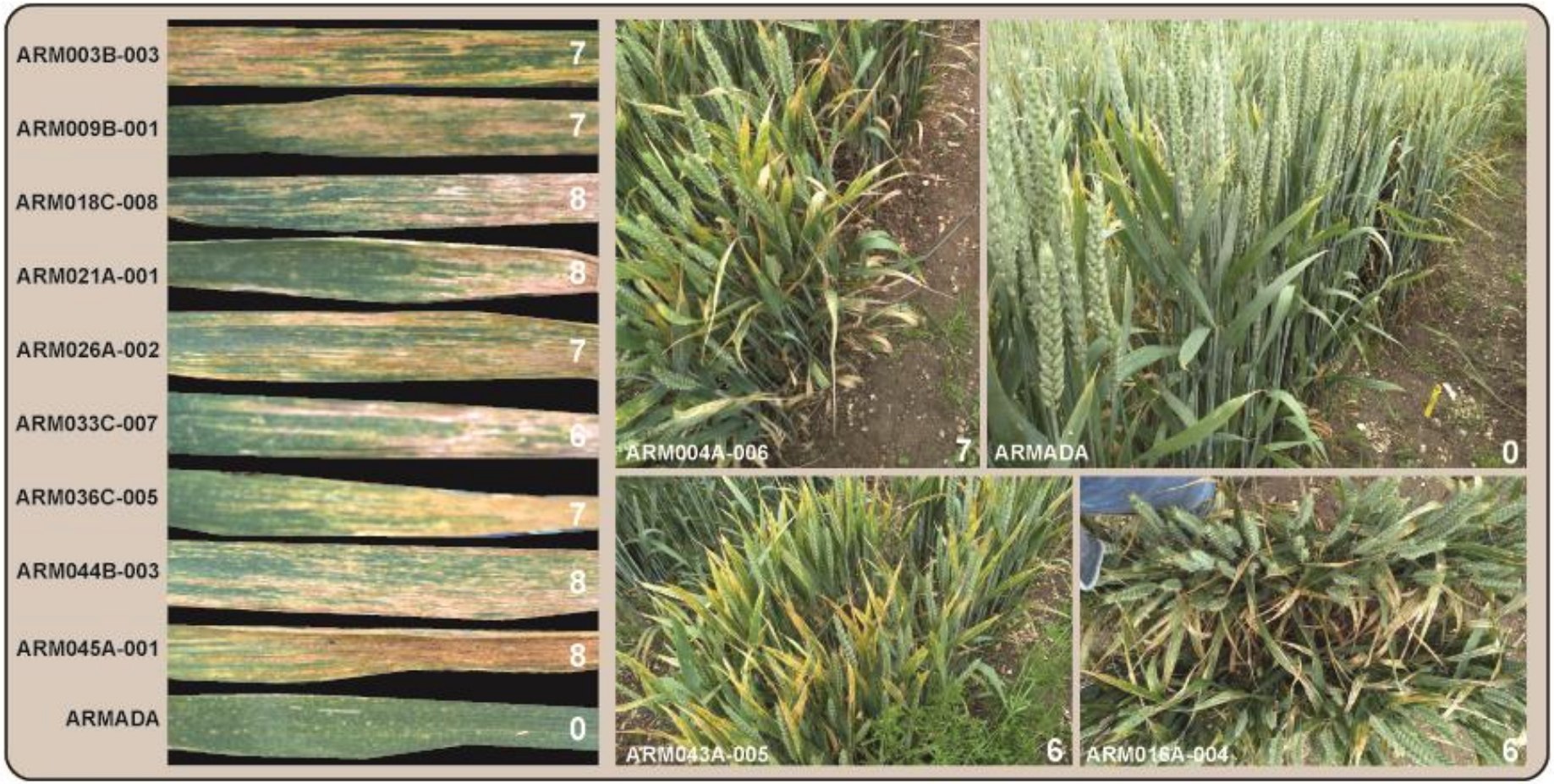


Figure 2-7. Yellow rust disease scores of the 12 selected Armada mutant lines for MutRenSeq.

2.4. Discussion

2.4.1. Number of susceptible mutant lines identified and confirmed in our EMS-based screen is relevant with what we observed in the literature

Forward genetic screens based on EMS-mutagenized populations have been successful in identifying relevant candidates for resistance genes in wheat. Mutants have been used to clone several genes: *Lr10*[170], *Tsn1*[171], *Sr22*/*Sr45*[137], *Sr33*[172], *Yr15*[146]. We hypothesized that *Yr7*, *Yr5* and *Yr12* encode NLR immune receptors and as such, a mutagenesis approach would be suitable to identify enough susceptible independent mutant lines to clone these genes using MutRenSeq.

We screened 1,000 mutant Cadenza lines for *Yr7* and successfully identified seven loss of function lines that were confirmed in M_3 and M_4 generations. For *Lr10*, the authors screened 52,000 M_2 seedlings and identified 33 susceptible lines from which three were confirmed in M_3 generation, although it is not clear whether the 52,000 seedlings included several replicate for the same line[170]. There is no information about how many mutants lines were screened to identify *Tsn1* mutants[171]. Authors screened 1,300 and 680 M_2 families and identified six susceptible mutants confirmed in M_3 generation in both populations and for *Sr22* and *Sr45*, respectively[137]. Two *Yr15* loss of function mutants out of 2,112 tested lines and eight out of 1,002 lines were identified in tetraploid and hexaploid accessions, respectively[146]. For *Sr33*, 850 M_2 families were screened and nine susceptible mutant lines were identified[172]. The frequency of observed susceptible plants in the Lemhi-Yr5 mutants was slightly higher than in the other examples (Table 2-10), although we will see in Chapter 3 that only one gene carries mutations in all the tested lines. Apart from *Lr10*, these numbers are similar to what we obtained for *Yr7*.

Table 2-10. Comparison of the number of loss of function mutants for a targeted resistance gene identified in EMS-mutagenesis screens in the literature.
Green highlight shows the two screens that were conducted in this thesis

Gene	#Lines screened	#Susceptible lines identified (frequency)
Lr10[170]	52,000	33 (0.06%)
Sr22[137]	1,300	6 (0.4 %)
Sr45[137]	680	6 (0.9%)
Yr15[146]	2,112 (tetraploid wheat)	2 (0.09%)
	1,002	8 (0.8%)
Sr33[172]	850	9 (1%)
Yr5[174]	500	7 (1.4%)
Yr7	1,000	7 (0.7%)
Yr12	1,451	17 (1.2%)

For *Yr12*, there is no *Pst* isolate able to discriminate for this gene only to our knowledge (Armada also carries *Yr3a* and *Yr4a*). Very little information is available on *Yr12* virulence given that it is not part of the differential set[163]. Thus, it was not possible to carry out glasshouse tests and we could only rely on field trials to select mutant lines having consistent susceptible phenotype through three successive years. Knowing that most of the current *Pst* isolates in the UK are virulent to *Yr3a* and *Yr4a* and including Armada controls in our field trials increased the likelihood of selecting for *Yr12* loss of function mutants.

We screened 1,451 M_2 families and selected the 12 most consistent susceptible lines. Five additional lines also had suitable criteria, bringing the total number to 17. This number seemed slightly higher than what we observed for *Yr7*, *Yr15* and *Sr22*/*Sr45*, but close to *Sr33* and *Yr5*'s suppressor screen. It could either be due to a different response from Armada, and probably the cultivar used to generate *Sr33* and *Yr5* mutants, to the EMS mutagenesis leading to a higher mutation density and thus less lines are needed to

recover the same number of lines carrying a mutation in the gene of interest, or more than one gene might be crucial for *Yr12*-mediated resistance. We mentioned in Chapter 1 that a sub-class of NLR work in pairs, with both partners being required to trigger resistance response. It could thus be that such a mechanism is driving *Yr12*-mediated resistance.

Results from our forward genetic screens are consistent with what was observed for similar approaches in the literature. We can thus be confident that, assuming *Yr7*, *Yr5* and *Yr12* encode dominant NLR immune receptors, MutRenSeq will be a suitable approach to identify candidates for these genes.

2.4.2. Summary

Using forward genetics involving EMS-mutagenesis, we successfully identified loss of function mutant for two targeted *Yr* genes in this work: *Yr7* and *Yr12*. The susceptible phenotype was confirmed in two successive progeny tests in the *Yr7* mutants and across three generations for the *Yr12* mutants. Additionally, we confirmed the phenotype of published *Yr5*- loss of function mutants in the Lemhi-Yr5 background. Given that the number of susceptible lines identified in our screens were consistent with what we observed in the literature, we are confident that MutRenSeq can be used to clone the corresponding genes.

3. *Yr7*, *Yr5* and *YrSP* encode BED-containing NLR proteins

3.1. Introduction

3.1.1. Marker-assisted selection to deploy resistance genes

The genetic relationship between *Yr5* and *Yr7* has been debated for almost 45 years[158,159]. Both genes map to chromosome arm 2BL in hexaploid wheat and were hypothesized to be allelic[154], and closely linked with *YrSP*[180]. We mentioned in Chapter 2 that *Yr7* originally comes from durum wheat cultivar 'Iumillo' and was introgressed into bread wheat cultivar 'Thatcher'. Thatcher is the main *Yr7* donor that we know of in modern bread wheat varieties. *Yr5* comes from spelt wheat 'Album' and among commercial lines carrying *Yr5* are several elite cultivars derived from the UC Davis breeding programme carrying a *Yr5+Yr15* introgression (https://dubcovskylab.ucdavis.edu/breeding). The same combination is used in Punjab Agricultural University breeding program, (Cristobal Uauy, personal communication) and likely in other breeding programs, although it is not necessarily documented. *YrSP* originates from bread wheat cultivar 'Spaldings Prolific' and, to our knowledge, it has not been widely deployed in commercial varieties. Whilst only two of > 6,000 tested *Pst* isolates worldwide have been found virulent to *Yr5*[155,157] (Table 2-1), both *Yr7* and *YrSP* have been overcome in the field. For *Yr7*, this is likely due to its wide deployment in cultivars (Figure 2-1). This highlights the importance of stewardship plans (including diagnostic markers) to deploy *Yr5* in combination with other genes as is currently being done with the *Yr5+Yr15* combination.

Marker-assisted selection, among other techniques described in Chapter 1, aims to facilitate selection of traits of interest for breeders (e.g. yield, quality, resilience to abiotic stress, disease resistance, etc …) based on the linkage between a marker (morphological,

biochemical or DNA/RNA variation) and the gene(s) involved in the expression of the traits[181,182]. The strength of the genetic linkage between the marker and the trait relies on how often both the marker and the expression of the phenotype of interest are found together in a given progeny. A gene-specific marker does not strictly have to be located within the allele of interest. It can be outside the locus, although it has to be genetically linked to the targeted allele to ensure that selecting the marker ensure selecting the right allele in 100 % of the cases. Alternatively, 'perfect marker' refers to a marker that is targeting the causal variation in the allele of interest. Both types of markers thus reduce the risk of false positive/negative, as no recombination can occur between the marker and the polymorphism associated with the trait of interest.

However, designing gene-specific markers or 'perfect markers' is not a trivial task. Indeed, one first needs to know exactly which gene(s) is crucial for the expression of the phenotype of interest. One then needs to either identify a polymorphism genetically linked to this variant of interest (e.g Simple Sequence Repeat marker for gene-specific marker) or identify variation among different alleles of this gene(s) to discriminate between the causative allele that is linked to the phenotype and other alleles for 'prefect marker'. Moreover, complex traits such as yield and quality often rely on the expression of multiple genes. This means that selecting only one specific variant of one gene may not have a strong effect on such complex traits. However, we saw in Chapter 1 how the development of SNP arrays coupled with phenotyping facilitated uncovering the genetic linkage between a wide range of markers and phenotype of interest. Breeders can then use this information to run a subset of these markers in their programs to select for a given phenotype[121,122,183,184].

In our case, *Yr7*, *Yr5* and *YrSP* are hypothesized to be single-dominant resistance genes (discussed in Chapter 2). Thus, we assumed only one gene is responsible for the corresponding resistant phenotype. Cloning *Yr7*, *Yr5* and *YrSP* would consequently enable development of gene-specific markers, or even 'perfect markers', that we described above. Consequently, the presence of the selected alleles would ensure the expression of the resistant phenotype. It is nevertheless important to note that this assumes that no other gene(s)/regulatory mechanisms could interfere with the expression of the resistance. Indeed, it has been reported in rice that resistance gene transfer between varieties does not always lead to expression of the resistance[185].

We showed in Chapter 2 how genetic mapping was previously used to determine the physical location of *Yr7*, *Yr5* and *YrSP* on chromosome arm 2BL. In this Chapter, we used two techniques in addition to genetic mapping to identify and validate candidate genes for *Yr7*, *Yr5* and *YrSP*: bulked segregant analysis coupled with exome capture and sequencing and mutational genomics coupled with Resistance gene enrichment Sequencing (MutRenSeq). We already presented and illustrated the principle of these three techniques in Chapter 1. However, because MutRenSeq is a more recent technique and our main focus in this work, we will provide more details on this approach in the following section.

3.1.2. Mutational genomics coupled with Resistance gene enrichment Sequencing (MutRenSeq).

3.1.2.1. Principle

MutRenSeq was developed during the course of this thesis (2016)[137] and Resistance gene enrichment Sequencing (RenSeq) itself was developed in 2013[52]. RenSeq is a targeted

sequencing technique for resistance genes strictly belonging to the NLR (nucleotide binding-site leucine-rich repeat) family. Targeting NLRs only is possible given the very characteristic and specific domain organisation in the derived proteins. We described the domain structure of these proteins in Chapter 1.

RenSeq was initially coupled with bulked segregant analysis to identify molecular markers that co-segregate with a pathogen resistance trait of interest[52]. Instead of investigating a segregating population for resistance and susceptibility, MutRenSeq relies on the development of mutagenised populations to identify loss of function mutants in the targeted gene (Figure 3-1).

Mutant population development (Figure 3-1, A; *Yr12* example in Figure 2-3)

The first step in MutRenSeq is to develop a mutant population in a cultivar carrying the resistance gene of interest. Preferably, the gene has been introgressed into an otherwise susceptible cultivar so all loss of resistance mutants are likely to carry a mutation in the gene of interest or a gene that is needed for the expression of the resistance mediated by the gene of interest (Figure 3-1, A). For example, this was the case for Lemhi-Yr5 and the AvocetS-Yr lines that we described in Chapter I. Alternatively, providing the pathogen isolate and the cultivar that are used are well characterised in terms of resistance genes and virulence/avirulence profile, it is possible to use other cultivars. This is what we did with Cadenza as we knew it carried *Yr7* and *Yr6* and we used a *Pst* isolate that could discriminate between these two genes in the pathology test (*Yr7* avirulent/ *Yr6* virulent). The susceptible phenotype then needs to be confirmed in the progeny of the identified loss of resistance mutant lines. This is to ensure that the observed susceptibility is real and not due to an experimental issue, such as disease escape or urediniospore contamination with a virulent *Pst* isolate.

This step is the main difference between the first RenSeq experiment and MutRenSeq. Indeed, we stated above that RenSeq was first coupled to bulk segregant analysis to fine-map resistance genes[52]. The resolution was not high-enough to identify the causal gene(s). Here used induced variation in a given wheat cultivar (EMS-mutagenised population). Comparing sequences derived from wild-type (resistant) and multiple independent mutant lines (susceptible) would thus enable identification of causal mutation(s) and subsequently causal gene(s). This is the principle of mutational genomics[177].

Given that many NLR genes are organised in clusters in the genome (concept first described by Michelmore and Meyers (1998)[186]), this close-proximity hinders identification of the causal gene among the others by genetic mapping. Indeed, the closer two genes are, the less likely a cross-over event will occur between them. Moreover, in wheat, it is known that long chromosomal fragments show suppressed recombination[187]. This means that even across large genomic regions, recombination rates can be low and thus the task of identifying causal gene by genetic mapping is even more difficult.

In this respect, the use of EMS-mutagenised population enabled us to circumvent the limitations of genetic mapping.

Bait library design and capture, enrichment and sequencing (Figure 3-1, B and C)

The next step in MutRenSeq involves extracting DNA from the selected independent mutant lines and the wild-type parent for the capture, enrichment and sequencing of NLR genes. We stated in Chapter 1 that NLRs show a very specific domain organisation and

this can be used to design baits for targeted sequencing. Initially, only the coding region was used to design the baits[51,52,137], although latest on-going improvements of the current wheat bait library now include intronic regions (Brande Wulff, personal communication). The baits are synthesized as RNA baits to improve the hybridisation with DNA[52]. In this thesis we used an improved version of the bait library initially used in the first MutRenSeq experiment[137]. This bait library was designed from published *Triticeae* genomes with associated gene annotation including *Triticum aestivum*, *Triticum durum*, *Aegilops sharonensis*, *Aegilops speltoides*, *Aegilops tauschii*, *Triticum urartu*, *Hordeum vulgare* and *Brachypodium distachyon*[137]. NLR Parser[188] was used to identify NLRs among coding sequences from annotated genes and RepeatMasker was used to eliminate repetitive regions in the genes to avoid designing baits that would hybridise to these sequences, as it would lead to capture of sequences that are unrelated to NLRs. Baits were designed as 120 nucleotide long sequences overlapping over 50 % of their sequences. Nearly identical baits were then removed to reach a final set of 60,000 sequences. Each of these sequences is ligated to biotinylated beads. It is important to note that the baits are able to bind a sequence that is at least > 80 % similar[189]. Thus, it is possible to capture NLRs (with at least 80% sequence similarity) that have not been annotated and that are not present in the bait library.

DNA samples from wild-type and mutant lines are mixed in solution with the biotinylated beads linked to the baits to allow hybridisation of the baits to NLR sequences. Streptavidin is used to pull-out the biotinylated beads and consequently any sequence that hybridized with the baits. This NLR-enriched sample is then sequenced with any sequencing technology. In our case we used HiSeq2500 (and MiSeq for Cadenza wild-type).

MutantHunter pipeline (Figure 3-1, C)

Sequencing data from the wild-type parent are used to produce a *de novo* assembly of the captured sequences and NLR-Parser allows identification of contigs showing NLR motifs. Sequencing data from the mutant lines are aligned onto the wild-type assembly, allowing SNP calling between the wild-type reference and the mutants. MutantHunter[137] processes the SNP calling files and parse contigs for which X number of mutant lines carry a mutation in the same contig. The more mutant lines which carry a mutation in a given contig, the more likely this contig is of being involved in the target phenotype[134] (discussed in Chapter 2).

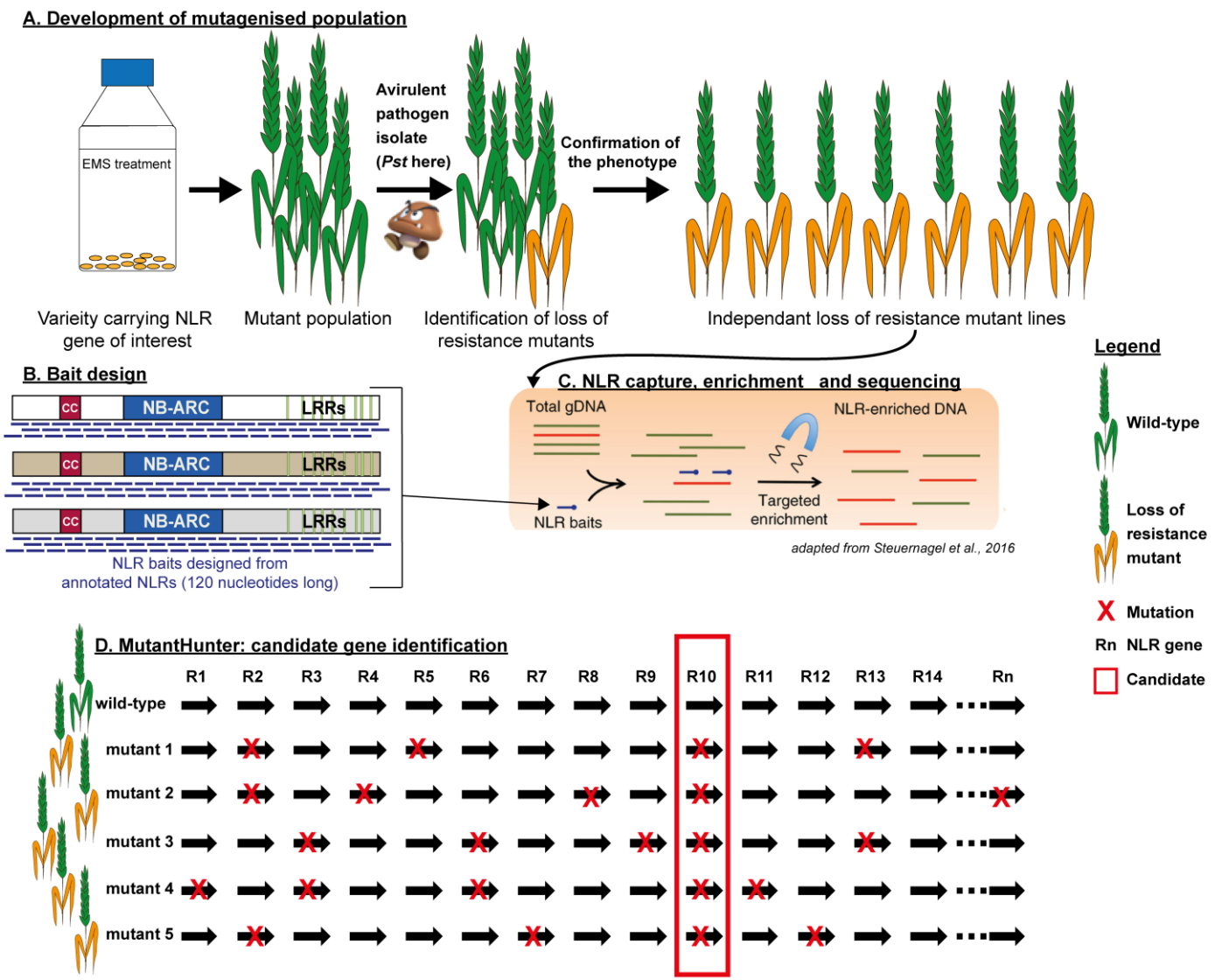


Figure 3-1. Illustration of MutRenSeq workflow

A. Development of the mutant population and selection of the loss of resistance lines. A detailed example for this step is described in Chapter 1.

B. Baits designed based on NLR sequences hybridize with NLRs in a genomic DNA sample

C. NLR capture enrichment and sequencing: Baits are biotinylated so they can be selected with streptavidin and the hybridized NLR sequences are thus separated from the DNA sample. Then NLR-enriched sample is then sequenced

D. MutantHunter is a program that is able to identify contigs that carry a mutation in a given number of mutant lines. On the illustration, each mutant line carries a mutation in NLR gene "R10". R10 is thus the best candidate in this example.

3.1.2.2. Limitations

MutRenSeq relies on the assumption that the targeted gene belongs to the NLR family. We mentioned in Chapter 1 that resistance genes do not systematically encode NLR proteins. For example, *Yr36* confers resistance to *Pst* and encodes a kinase with a putative lipid-binding domain[63]. It is thus mandatory to have strong evidence for a given gene of interest to encode an NLR protein before carrying out a MutRenSeq approach, as it will not be captured if that is not the case.

We discussed above that mutational genomics does not rely on using segregating populations and thus enabled us to overcome the issue of suppressed recombination that can occur across large chromosomal regions in wheat. However, it is important to bear in mind that developing a mutagenised population still requires time and confirmation of the phenotype in the next mutant generation is highly advised. Indeed, if a given mutant line is not a true susceptible line, it will greatly reduce the power of the analysis and consequently the positive signal will be diluted.

NLRs mostly mediate seedling resistance and it can thus be tested very early on during the development of the plant (one leaf stage, Zadok's stage 11), which is convenient for selecting early on which plants to extract DNA from. However, seedling tests require a *Pst* isolate able to specifically induce a resistant/susceptible response when the targeted gene is functional/non-functional, respectively. Although these conditions are often met, it is not always the case, as for *Yr12*. Indeed, *Yr12* is an Adult Plant Resistance (APR) gene for which we do not have a specific *Pst* isolate able to differentiate between *Yr12* and other resistance genes. Therefore, in this thesis, we had to test the *Yr12* susceptible Armada lines in the field across several years (Chapter 2).

Identifying candidate genes involved in disease resistance with MutRenSeq would be more difficult if several genes are involved. Indeed, in this case the identified susceptible mutants would carry mutation in different genes. The MutantHunter analysis would thus have to be adapted. For example, the settings should be set to identify contigs that carry mutation in at least half of the susceptible mutant lines, if two genes are involved. It is thus helpful to carry out complementation tests in the identified mutant lines to be able to adapt the analysis accordingly, which means going through more generations and pathology tests.

3.1.2.3. Suitability to use MutRenSeq for cloning Yr7, Yr5 and YrSP

We hypothesized that MutRenSeq represents the most suitable option to clone *Yr7*, *Yr5*, *YrSP*, despite some of the limitation listed above. First of all, we know that all three genes are race-specific resistance genes controlling seedling resistance. Thus, pathology tests can be carried out very early on in plant development. Second, both varieties/near isogenic lines carrying the genes and *Pst* isolates with suitable virulence profiles to specifically identify functional and non-functional resistance are well characterised. Therefore, seedling tests can be carried in controlled-environment cabinets. From the different mapping studies and allelism tests presented in Chapter 2, there is strong evidence for all three genes to be single dominant. This means that only one gene is likely to control the resistance. Thus, only one homozygous mutation should be sufficient to lead to a loss of function phenotype. Altogether, this suggests that MutRenSeq is a suitable solution to clone *Yr7*, *Yr5* and *YrSP*.

3.1.3. Summary and Disclaimer

In this Chapter we will present how we used a MutRenSeq approach to clone *Yr7*, *Yr5* and *YrSP*. We validated these candidate genes with genetic mapping, bulked segregant analysis followed by exome capture enrichment sequencing and presence/absence of the candidates in varieties known to carry *Yr7*, *Yr5* and *YrSP*. We developed diagnostic markers to allow for selection for *Yr7*, *Yr5* and *YrSP* in breeding programs.

All contents of this chapter are published in Marchal et al., 2018[153], except from the N-terminus structure prediction analysis in Yr7 and Yr5 proteins. Several experiments were carried out as part of a collaboration with Evans Lagudah (EL) and Jianping Zhang (JZ) (CSIRO), Robert McIntosh (RM) and Peng Zhang (PZ) (University of Sydney), Paul Fenwick (PF) and Simon Berry (SB) (Group Limagrain UK). Table 3-1 below shows the specific experiments carried out by our collaborators and all other presented experiments were carried out by myself under the supervision of Cristobal Uauy, Brande Wulff and Simon Berry.

Table 3-1. Contributions of our collaborators to the work presented in this Chapter

Experiment	Contributed by
Generation of F_2 populations derived from a cross between AvocetS-Yr5 and AvocetS; AvocetS-YrSP and AvocetS to genetically map *Yr5* and *YrSP*, respectively	SB, PF
Generation of F_2 populations derived from a cross between AvocetS-Yr5 and AvocetS-YrSP; AvocetS-Yr7 and AvocetS-YrSP to determine whether *YrSP* is allelic to *Yr5* and *Yr7*	RM, PZ
MutRenSeq on AvocetS-YrSP mutants	El, JZ
Confirmed the start ATG and stop codon with 5' and 3' Rapid Amplification of cDNA Ends (RACE) PCR in *Yr5* and *Yr7*	JZ

3.2. Materials and methods

3.2.1. Plant materials and *Pst* isolates used to clone *Yr7*, *Yr5* and *YrSP*

Table 3-2 and Table 3-3 summarise the *Pst* isolates and different plant resources used in this Chapter, respectively, and we will provide additional details in this section. We described *Yr7* and *Yr5* loss of function mutants in Cadenza and Lemhi-Yr5 and the Avocet-*Yr5*/*Yr7*/*YrSP* mutants in Chapter 2. Additionally, we developed two F_2 populations based on a cross between the susceptible mutant line Cad127 to the Cadenza wild-type (139 individuals) and between Cad1978 and Cadenza wild-type (192 individuals). The aim of these populations was to assess the genetic linkage between the *Yr7* candidate gene arising from the MutRenSeq analysis and the *Yr7* locus via traditional genetic mapping and bulked segregant analysis. We investigated two additional F_2 populations between AvocetS and the NILs carrying the corresponding *Yr* gene (376 individuals for *Yr5* and 94 for *YrSP*) to map *Yr5* and *YrSP* candidates.

We explored four different wheat panels to determine *Yr7*, *Yr5* and *YrSP* prevalence in breeding materials: a core set of the Watkins collection, which represent a set of global bread wheat landraces collected in the 1920-30s[190], the Gediflux collection that includes modern European bread wheat varieties (1920-2010)[191], a set of varieties that belonged to the UK AHDB Recommended List (https://cereals.ahdb.org.uk/varieties/ahdb-recommended-lists.aspx) between 2005 and 2018 and a bespoke set of putative *Yr7* carriers based on a literature search.

Table 3-2. Virulence profiles of the *Pst* isolates we used in this study.
†Virulence profile determined with the Johnson et al., (1972) nomenclature[163]

Pst isolate	Virulence profile	Reference
PST 15/151	*Yr1, Yr2, Yr3a, Yr3b, Yr4b, Yr6, Yr9. Yr17, YrSd, Yr32, YrRo, YrSol*	UKCPVS report 2016
PST 14/106	*Yr1, Yr2, Yr3a, Yr4, Yr6, Yr7, Yr9, Yr17, Yr25, Yr32, YrSp, YrRo, YrSo*	UKCPVS report 2015
PST 08/21	*Yr1, Yr2, Yr3, Yr4, Yr6, Yr9, Yr17, Yr32, YrRob, and YrSol*	Hubbard et al., 2015[173]
PST 81/20	*Yr2, Yr3, Yr4, Yr6*	McGrann *et al.*, 2014[174]
108 E141A+	*Yr2, Yr3, Yr4, Yr6, YrSd, YrSp*†	University of Sydney Plant Breeding Institute Culture no. 420
150 E16A+	*Yr6, Yr7, Yr8, Yr9, Yr10*†	University of Sydney Plant Breeding Institute Culture no. 598
134 E16A+	*Yr6, Yr7, Yr8, Yr9*†	University of Sydney Plant Breeding Institute Culture no. 572

Table 3-3. Plant materials analysed in the present Chapter and corresponding *Pst* isolates used for the pathology assays.
For *Pst* strain virulence profiles see Table 3-2

Gene	Experiment	Plant Material	Rust isolate	Reference(s)
Yr7	MutRenSeq	EMS-derived TILLING population in the UK Cadenza cultivar	PST 08/21	Krasileva et al., 2017
	Confirmation of the *Yr7* candidate through sequencing	AvocetS-Yr7 EMS mutants	108 E141 A+ (University of Sydney PBI Culture no. 420)	Generated for the study
	Genetic linkage confirmation	F_2 population: Cad127 x CadWT (139) F_2 population: Cad1978 x CadWT (192)		Generated for the study
	Yr7 KASP primer testing	Cadenza-derived varieties + Yr7 carriers	PST 08/21; PST 15/151; PST 14/106	Generated for the study
	Yr7 frequency in breeding materials	Set of varieties from the UK Recommended list between 2005 and 2018 (AHDB)		https://cereals.ahdb.org.uk/varieties/ahdb-recommended-lists.aspx
		Gediflux collection		Reeves et al., 2004
		Core-set of the Watkins collection		Wingen et al., 2014
Yr5	MutRenSeq	EMS-derived Lemhi-Yr5 mutants	PST 81/20	McGrann et al., 2014
	Confirmation of the *Yr5* candidate through sequencing	AvocetS-Yr5 EMS mutants	150 E16 A+ (University of Sydney PBI Culture no. 598)	Generated for the study
	Genetic linkage confirmation	F_2 population: AvocetS x AvocetS-Yr5 (376)		Generated for the study
YrSP	MutRenSeq	AvocetS-YrSP EMS mutants	134 E16 A+ (University of Sydney PBI Culture no. 572)	Generated for the study
	Genetic linkage confirmation	F_2 population: AvocetS x AvocetS-Yr5 (94)		Generated for the study

3.2.2. DNA preparation and Resistance gene enrichment Sequencing (RensSeq)

We extracted total genomic DNA from young leaf tissue using the large-scale DNA extraction protocol from the McCouch Lab (https://ricelab.plbr.cornell.edu/dna_extraction) and a previously described method[192]. We checked DNA quality and quantity on a 0.8 % agarose gel and with a NanoDrop spectrophotometer (Thermo Scientific). Arbor Biosciences (Ann Arbor, MI, USA) performed the targeted enrichment of NLRs according to the MYbaits protocol using Triticeae RenSeq Bait Library V2 (sequences available at https://github.com/steuernb/MutantHunter). Library construction was performed using the TruSeq RNA protocol v2 (Illumina 15026495). Libraries were pooled with one pool of samples for Cadenza (*Yr7*) mutants and one pool of eight samples for the Lemhi-Yr5 parent and Lemhi-Yr5 mutants. AvocetS-Yr5 and AvocetS-YrSP wild-type, together with their respective mutants, were also processed according to the MYbaits protocol and the same bait library was used. All enriched libraries were sequenced on a HiSeq 2500 (Illumina) in High Output mode using 250 bp paired end reads and SBS chemistry. For Cadenza wild-type, we generated data on an Illumina MiSeq instrument. In addition to the mutants, we also generated RenSeq data for tetraploid Kronos and hexaploid Paragon to look for *Yr5* and *Yr7* alleles in these cultivars. Details of all the lines sequenced, alongside NCBI accession numbers, are presented in Appendix 8-4 and Appendix 8-5.

3.2.3. MutantHunter pipeline

We adapted the pipeline from https://github.com/steuernb/MutantHunter/ to identify candidate contigs for the targeted *Yr* genes. First, we trimmed the RenSeq-derived reads

with trimmomatic[193] using the following parameters: ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 MINLEN:50 (v0.33). We made *de novo* assemblies of wild-type plant trimmed reads with the CLC assembly cell and default parameters apart from the word size (-w) parameter that we set to 64 (v5.0, http://www.clcbio.com/products/clc-assembly-cell/) (Table 3-4)

Table 3-4. *de novo* assemblies generated from the corresponding RenSeq data. Complete NLRs were defined as carrying both NB-ARC and LRR motifs

***de novo* assembly**	**assembler**	**#contigs**	**#NLR-contigs**	**#complete_NLRs**
Cadenza-WT	CLC assembly cell	29706	5572	431
Lemhi-Yr5	CLC assembly cell	352145	5174	862
AvocetS	CLC assembly cell	400158	5574	829
AvocetS-YrSP	CLC assembly cell	530695	5341	904
AvocetS-Yr7	CLC assembly cell	278126	5299	887
AvocetS-Yr5	CLC assembly cell	362856	5355	863
Paragon	CLC assembly cell	2571400	4744	494
Kronos	CLC assembly cell	255977	3651	516
AvocetS-YrSP-WT	CLC Genomics Workbench	268235	5361	791
AvocetS-Yr5-WT	CLC Genomics Workbench	109608	5180	782

To test whether CLC assembly cell was a suitable to assemble contiguous NLR contigs, we tested the Masurca assembler[194] in parallel (v3.3.1). We used the default parameters with an insert size of 700 bp and a standard deviation of 200 based on the library quality check we received from Novogene. We then compared the two assemblies based on the number of complete NLRs assembled (carrying both NB-ARC and LRR motifs) and observed that overall CLC assemblies performed better (Table 3-5). We thus focused on these assemblies for the following steps.

Yr7 mutant analysis:

For Cadenza mutants, we used the following MutantHunter program parameters to identify candidate contigs: -c 20 -n 4-7 -z 1000. These options require a minimum coverage of 20x for SNPs to be called; we did several runs with -n varying from 4 to 7, -n 4 means that at least four susceptible mutants must have a mutation in the same contig to report it as candidate; small deletions were filtered out by setting the number of coherent positions with zero coverage to call a deletion mutant at 1000. We used Cadenza genome assembly available from the Earlham Institute (Appendix 8-6, http://opendata.earlham.ac.uk/Triticum_aestivum/EI/v1.1) with the NLR-Annotator program using default parameters[195] (https://github.com/steuernb/NLR-Annotator) to reconstruct the full-length *Yr7* candidate.

Yr5 and *YrSP* mutant analysis:

Regarding the Lemhi-Yr5 mutants used to clone *Yr5*, we used the same MutantHunter parameters as described above for Cadenza mutants (*Yr7*). To identify *Yr5* and *YrSP* contigs from Avocet mutants, we followed the MutantHunter pipeline with all default parameters, except in the use of CLC Genomics Workbench (v10) for reads QC, trimming, *de novo* assembly of Avocet wild-type and mapping all the reads against *de novo* wild-type assembly. Default MutantHunter parameters were used except that –z was set as 100. The parameter –n was set to 2 in the first run and then to 3 in the second run.

3.2.4. Sequence confirmation of the candidate contigs and gene annotation

We sequenced the *Yr7*, *Yr5*, and *YrSP* candidate contigs from the mutant lines to confirm the EMS-derived mutations using primers documented in Appendix 8-7. We first PCR-amplified the complete locus from the same DNA preparations as the ones submitted for RenSeq with the Phusion® High-Fidelity DNA Polymerase (New England Biolabs)

following the suppliers protocol (https://www.neb.com/protocols/0001/01/01/pcr-protocol-m0530). We then carried out nested PCR on the obtained product to generate overlapping 600-1,000 bp amplicons that were purified using the MiniElute kit (Qiagen). The purified PCR products were sequenced by GATC following the LightRun protocol (https://www.gatc-biotech.com/shop/en/lightrun-tube-barcode.html). Resulting sequences were aligned to the wild-type contig using ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/). This allowed us to curate the *Yr7* locus in the Cadenza assembly that contained two sets of unknown ('N') bases in its sequence, corresponding to a 39 bp insertion and a 129 bp deletion, and to confirm the presence of the mutations in each mutant line (Figure 3-3).

We used HISAT2[196] (v2.1) to map RNA-Seq reads available from Cadenza and AvocetS-*Yr5*[197] to the RenSeq *de novo* assemblies with curated loci to define the structure of the genes (Figure 3-12). We used the following parameters: --no-mixed --no-discordant to map reads in pairs only. We used the --novel-splicesite-outfile to predict splicing sites that we manually scrutinised with the genome visualisation tool IGV[198] (v2.3.79). Predicted coding sequences (CDS) were translated using the ExPASy online tool (https://web.expasy.org/translate/). This allowed us to predict the effect of the mutations on each candidate transcript (Figure 3-6, Appendix 8-4). The long-range primers for both *Yr7* and *Yr5* loci were then used on the corresponding susceptible Avocet NIL mutants to determine whether the genes were present and carried mutations in that background (Appendix 8-4).

To determine whether *Yr7*, *Yr5*, and *YrSP* encode Coiled Coil (CC) domains we used the NCOILS prediction program[199] (v1.0, https://embnet.vital-it.ch/software/COILS_form.html) with the following parameters: MTIK matrix with applying a 2.5-fold weighting of positions a, d (Figure 3-10). Additionally, we used the

webtool Phyre2[200] (v2.0 http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) to predict the structure of the Yr7 amino-acid sequence from the start residue to the beginning of the BED domain to determine whether it had homology with existing Coiled Coil containing proteins (Figure 3-11).

3.2.5. Genetic linkage confirmation

We used a set of F_2 populations to genetically map the candidate contigs (Table 3-3). We extracted DNA from leaf tissue at the seedling stage (Zadok's scale 11) following a previously published protocol[201] and Kompetitive Allele Specific PCR (KASP) assays were carried out as previously described[132]. We used R/qtl package[202] to generate the genetic map based on a general likelihood ratio test and genetic distances were calculated from recombination frequencies (v1.41-6).

We used previously published markers linked to *Yr7*, *Yr5*, and *YrSP* (WMS526, WMS501 and WMC175, WMC332, respectively[180,203,204]) in addition to closely linked markers WMS120, WMS191, and WMC360 (based on the GrainGenes database https://wheat.pw.usda.gov/GG3/) to define the physical *Yr* locus on the Chinese Spring assembly RefSeq v1.0[112]. We used two different approaches for genetic mapping depending on the material. For *Yr7*, we used the public data[107] for Cad127 (www.wheat-tilling.com) to identify nine EMS-induced mutations located within the *Yr7* physical interval based on BLAST analysis against RefSeq v1.0. We used KASP primers when available and manually designed additional ones including an assay targeting the Cad127 mutation in the *Yr7* candidate contig (Appendix 8-7). We genotyped the Cad127 F_2 populations using these nine KASP assays and confirmed genetic linkage between the Cad127 *Yr7* candidate mutation and the nine mutations across the physical interval (Figure 3-8).

For *Yr5* and *YrSP*, we first aligned the candidate contigs to the best BLAST hit in an AvocetS RenSeq *de novo* assembly (Table 3-4). We then designed KASP primers targeting polymorphisms between these sequences and used them to genotype the corresponding bi-parental F_2 population (Appendix 8-7). For both candidate contigs we confirmed genetic linkage with the previously published genetic intervals for these *Yr* genes (Figure 3-8).

3.2.6. Exome capture and sequencing in Cad1978

We screened 192 individuals derived from Cad1978 x Cadenza wild-type cross with PST 08/21 to determine the segregation ratio of resistant:susceptible phenotypes (Table 3-8). Inoculation method was the same as described in Chapter 2. We assembled bulks with equal amount of leaf tissue from 20 resistant and 20 susceptible individuals and extracted DNA from these two bulks as described in Section 3.2.2. Exome capture and sequencing was performed on the bulks by the Earlham Institute with the same array and protocol that was used for the Cadenza TILLING population[107].

We aligned the reads from Cadenza wild-type, Cad1978, CadWT x Cad1978 susceptible bulk and CadWT x Cad1978 resistant bulk to RefSeqv1.0 pseudomolecule parts[112] with bwa aln (v0.7.115)[205] and the samtools suite (v1.3.1). We used Freebayes[206] (v1.1.0) with default parameters to identify SNPs. Freebayes has an *in-built* quality filtering step when using default parameters so no subsequent filtering on quality was made. We used Cadenza wild-type data to filter-out any varietal SNPs between Cadenza and Chinese Spring (RefSeqv1.0) in the bulks and keep only the SNPs between Cadenza wild-type and Cad1978, which are hypothetical EMS-induced SNPs. These EMS-induced SNP positions were used to extract the corresponding information in the bulks. We applied filters on the depth of coverage (DP > 5), the number of reads supporting a given SNP

position, and calculated the allele frequencies as follow: the number of high-quality reads carrying the alternative allele (AO) divided by the total number of high-quality bases (DP). We plotted the allele frequencies across the chromosome and the results are shown in Figure 3-7.

3.2.7. Identification of *Yr7* and *Yr5* related sequences in sequenced wheat cultivars

We used the *Yr7* and *Yr5* sequences to retrieve the best BLAST hits in the *T. aestivum* and *T. turgidum* wheat genomes listed in Appendix 8-6. We reanalyzed RNA-Seq data from cultivar Kronos[207] to determine whether the Kronos *Yr5* allele was expressed. We followed the same strategy as that described to define the *Yr7* and *Yr5* gene structures (section 3.2.4). We generated a *de novo* assembly of the Kronos NLR repertoire from Kronos RenSeq data (Table 3-4) and used it as a reference to map read sequences derived from one replicate of wild-type Kronos at heading stage. Read depths up to 30x were present for the Kronos *Yr5* allele which allowed confirmation of its expression (Figure 3-12). Likewise, the RNA-Seq reads confirmed the gene structure, which is similar to *YrSP*, and the premature termination codon in Kronos *Yr5* (Figure 3-12). Whether this allele confers resistance against *Pst* remains to be elucidated.

3.2.8. Development and testing of diagnostic markers for *Yr7*, *Yr5* and *YrSP*

3.2.8.1. Yr7 gene specific markers design and testing

We aligned the *Yr7* sequence with the best BLAST hits in the genomes listed in Appendix 8-6 and designed KASP primers targeting polymorphisms that were Yr7-specific. We designed 54 primer sets in total and tested them on DNA from a subset of 96 hexaploid

wheat accessions that are part of the WAGTAIL panel (developed for the BBSRC LINK project "Wheat Association Genetics for Trait Analysis and Improved Lineages" (BB/J002607/1)), provided by Simon Berry (Group Limagrain). This subset of the panel contained four varieties that likely carry *Yr7* based on their pedigree (Tonic, Brock, Tommy, Grafton). Three markers Yr7-A, Yr7-B and Yr7-D were retained after testing (Figure 3-14).

We further tested the three markers (Yr7-A, Yr7-B, Yr7-D) on a selected panel of Cadenza-derivatives and cultivars that were positive for *Yr7* markers in the literature, including the *Yr7* reference cultivar Lee (Table 3-11, Appendix 8-9, Table 3-12). Paul Fenwick (Group Limagrain) screened the panel of Cadenza-derivatives with three Pst isolates: PST 08/21 (Yr7-avirulent), PST 15/151 (Yr7-avirulent) and PST 14/106 (Yr7-virulent) (Table 3-2) to determine whether the cultivars that were positive for *Yr7* also showed the expected IT for the presence of *Yr7* (Table 3-11). Pathology assays were performed as for the screening of the Cadenza mutant population. We retrieved pedigree information for the analyzed cultivars from the Genetic Resources Information System for Wheat and Triticale database (GRIS, www.wheatpedigree.net) and used the Helium software[208] (v1.17) to illustrate the breeding history of *Yr7* in the UK (Figure 3-15).

We used the three *Yr7* KASP markers to genotype (i) cultivars from the AHDB Wheat Recommended List from 2005-2018 (https://cereals.ahdb.org.uk/varieties/ahdb-recommendedlists.aspx); (ii) the Gediflux collection of European bread wheat cultivars released between 1920 and 2010[5]; and (iii) the core Watkins collection (3.2.1). KASP assays were carried out as described in 3.2.5 and results are reported in Table 3-11 and Appendix 8-9.

3.2.8.2. Yr5 and YrSP gene specific markers

We identified a 774 bp insertion in the *Yr5* allele 29 bp upstream of the STOP codon with respect to the Cadenza and Claire alleles. Genomic DNA from *YrSP* confirmed that the insertion was specific to *Yr5*. We used this polymorphism to design KASP primers tagging the insertion (GenBank #MN273772 and Appendix 8-8). Figure 3-16 describes how the three primers were designed.

We tested the primers on a set of cultivars listed in Table 3-12, including *Triticum aestivum* ssp. *spelta* cv. Album (*Yr5* donor) and bread wheat cultivar Spaldings Prolific (*YrSP* donor). The lack of amplification in some cultivars most likely represents the absence of the loci in the tested cultivars. For *YrSP*, we aligned the *YrSP* and *Yr5* sequences to design KASP primers targeting the G to C SNP between the two alleles (Figure 3-17, Appendix 8-8). We tested the marker by genotyping selected cultivars as controls and cultivars from the AHDB Wheat Recommended List from 2005- 2018 (Appendix 8-9).

3.2.9. Data availability

All sequencing data has been deposited in the NCBI Short Reads Archive under accession numbers listed in Appendix 8-5 (SRP139043). Cadenza (*Yr7*) and Lemhi (*Yr5*) mutants are available through the JIC Germplasm Resource Unit (www.seedstor.ac.uk). *Yr7*, *Yr5* and *YrSP* sequences (gDNA, CDS, protein) with mutation variants were deposited on GenBank (*Yr7*: MN273771.1, *Yr5*: MN273772, *YrSP*: MN273773).

3.3. Results

To clone the genes encoding *Yr7*, *Yr5*, and *YrSP*, we identified susceptible ethyl methanesulfonate-derived (EMS) mutants from different genetic backgrounds carrying these genes (presented in Chapter 2). We used seven independent Cadenza mutant lines to clone *Yr7* and seven Lemhi-Yr5 mutants to clone *Yr5*. In addition, our collaborators conducted an independent MutRenSeq analysis on four Avocet-Yr5 mutant lines to confirm our *Yr5* candidate and on four Avocet-YrSP lines to clone *YrSP*. All the results presented in this Chapter were published in Marchal et al., 2018[153], except for the structure prediction analysis.

3.3.1. Confirmation of NLR-enriched sequences in *Yr7* and *Yr5* RenSeq data

We compared assemblies generated with Masurca[194] and CLC Genomics Workbench for both *Yr7* and *Yr5* RenSeq datasets to select the one containing the most complete NLRs (Table 3-5).

Table 3-5. Comparison of assemblies derived from Masurca and CLC genomics workbench

	Yr7			***Yr5***	
	Masurca	CLC	Cadenza assembly	Masurca	CLC
#contigs	70253	121095	N/A	66984	352145
#Total NLRs	5573	4263	3266	5708	5175
#Complete NLRs	432	420	1438	617	862
#LRR-only	2084	1354	19	2322	1168

Comparable NLR numbers were identified in both Masurca and CLC assemblies for *Yr5* data and 1,000 more NLRs were identified in Masurca assembly in *Yr7* data. However, almost twice as many 'LRR-only' NLRs were identified in Masurca assemblies as

compared to CLC. Overall, CLC performed better than Masurca for the *Yr5* data but very similar number of complete NLRs were found in both assemblies for the *Yr7* data. Discrepancies between the *Yr7* and *Yr5* data could be due to different sequencing technologies used: MiSeq for Cadenza wild-type (*Yr7*) and HiSeq 2500 for Lemhi-Yr5.

A full-genome assembly for Cadenza was available at the time of the study (Appendix 8-6) and we assumed its completeness would be better than the *de novo* assembly of Cadenza NLRs from RenSeq data. We thus used NLR Annotator to identify putative NLR loci on this assembly and extended their boundaries by 3,000 bp on both 5 and 3' ends. To maximise our chances of having all Cadenza NLRs represented in the MutantHunter analysis, we used both *de novo* assembled and newly annotated and extended NLR loci as references to map Cadenza-derived RenSeq reads. Because CLC performed better than Masurca on *Yr5* data, we decided to use CLC-derived Lemhi-Yr5 assembly to clone *Yr5*.

Table 3-6 shows that coverage of NLRs was higher than the average coverage of all contigs for all samples in *Yr5* and *Yr7 de novo* assembly. This confirms that samples were enriched in NLR sequences and the associated average coverage is high enough to call mutations with confidence (66 to 206x).

Table 3-6. Average coverage per contig for all contigs and NLRs only in *Yr5* and *Yr7* datasets.

Targeted Gene	Line	Average coverage all contigs	Average coverage NLRs
Yr5	Lemhi-Yr5	6.4	140.9
	Lem115	4.6	135.8
	Lem241	4.7	129.6
	Lem287	5.8	168
	Lem387	6.6	202.4
	Lem474	4.7	119.6
	Lem500	6.1	206.3
	Lem095	7.4	180.4
Yr7	Cadenza	8.6	66.2
	Cad127	28.8	170.4
	Cad1551	15.8	113.3
	Cad1978	21.3	151.2
	Cad1034	20.6	149.2
	Cad855	20.8	164.6
	Cad903	14.1	113.1
	Cad923	20.7	158.1

3.3.2. Candidate gene identification for *Yr7* and *Yr5* with MutantHunter

The MutantHunter pipeline was run on the alignment files to identify NLRs for which '*X*' mutant lines carry a mutation. Table 3-7 illustrates the different runs with '*X*' varying between seven (all mutant lines carry the mutation) and four and allowing for up to two mutants to share the same mutation. Other parameters were kept as described in the corresponding Methods section (3.2.3).

None of the contigs seemed to have mutations in all seven mutant lines for *Yr7* or *Yr5*. However, allowing two mutant lines to share a mutation (Table 3-7, run E) identified the same contig as in run B for *Yr5*. Lem387 and Lem241 are thus likely to be sibling lines (Figure 3-4). For *Yr7* we identified a single contig carrying mutations in six out of seven mutant lines in alignments derived from the *de novo* RenSeq assembly. We did not identify mutations in this particular contig in line Cad903 (Table 3-7).

Table 3-7. Descriptions of MutantHunter results using different parameters.

	MutantHunter runs	**A**	**B**	**C**	**D**	**E**
	#Mutant lines allowed to share a common mutation	0	0	0	0	2
	Min #mutants having a mutation in the same contig	7	6	5	4	7
Yr7	#candidates contigs *Yr7* (*de novo* RenSeq assembly)	0	1	1	5	
	#candidates contigs *Yr7* (annotated Cadenza NLRs as assembly)	1[1]	1	1	5	
Yr5	#candidates contigs *Yr5* (*de novo* RenSeq assembly)	0	1	1	1	1[2]

[1] The contig found here in 100% identical to the one found in the *de novo* assembly (run B), only it is longer on both 5' and 3'ends. We identify a mutation in Cad903 in this contig, thus all the *Yr7* loss of function mutant we identified in Chapter 2 carry a mutation in this contig.

[2] This is the same contig as the one found in *Yr5* run B

In summary, after having identified candidate contigs for *Yr7* in Cadenza mutants and for *Yr5* in Lemhi-Yr5 mutants, we investigated them in more details in the other materials described in Table 3-3. Collaborator contributions regarding these experiments are documented in Table 3-1:

- *Yr7*: we first determined whether running MutantHunter on the annotated Cadenza NLRs would identify the same contig as in the *de novo* assembly for RenSeq data (see details below). Additionally, we investigated three AvocetS-*Yr7* susceptible mutant lines to determine whether they carried mutations in the same contig (Appendix 8-4)

- *Yr5*: an independent MutRenSeq experiment was carried out on four AvocetS-*Yr5* susceptible mutant to determine whether the same *Yr5* candidate contig would be identified.

- *YrSP*: we performed MutRenSeq on four AvocetS-*YrSP* susceptible mutant to identify candidate contigs for *YrSP*.

3.3.2.1. *Yr7 candidate*

We identified a candidate contig carrying a mutation in all mutant lines except Cad903. We thus hypothesized that either Cad903 carried a mutation in another gene that is important for *Yr7*-mediated resistance, or that the identified candidate contig was truncated and thus we could not observe a putative Cad903 mutation outside the assembled contig. This was documented in the first published MutRenSeq experiment, where the *Sr22* gene corresponded to two RenSeq contigs[137].

To determine which of these two hypotheses was the most plausible, we aligned the RenSeq data from Cadenza and the mutants to annotated NLRs from the Cadenza genome assembly and ran MutantHunter with the same parameters as above. We identified an NLR contig which had independent mutations in all seven mutant lines. This NLR was identical to the contig found in run B but was longer in both the 5' and 3' ends (Figure 3-2). The extended Cadenza assembly allowed us to identify a mutation in Cad903, the only mutant for which we had failed to previously identify a mutation in our top-ranking candidate contig. Thus, the Cad903 mutation was not retrieved previously because the corresponding NLR was incomplete in the *de novo* assembly, confirming our initial hypothesis.

There were two unknown nucleotide positions in the NLR sequence from the Cadenza assembly (each marked by a single N) that we corrected using both the RenSeq data and Sanger sequencing (Figure 3-3). These single N's corresponded to a 39 bp insertion and the deletion of 129 bp in the corrected sequence below.

Our collaborator Jianping Zhang (CSIRO), investigated this candidate in the three AvocetS-*Yr7* susceptible mutant lines provided by Peng Zhang and Robert McIntosh (University of Sydney). All lines carried a putative EMS mutation in the contig and both AvSYr7_2 and AvSYr7_3 carried the same mutation (Appendix 8-4, Figure 3-6). The candidate was thus confirmed in an independent mutant background.

We thus identified a candidate contig for *Yr7* that is supported by 10 independent mutant lines coming from two different backgrounds.

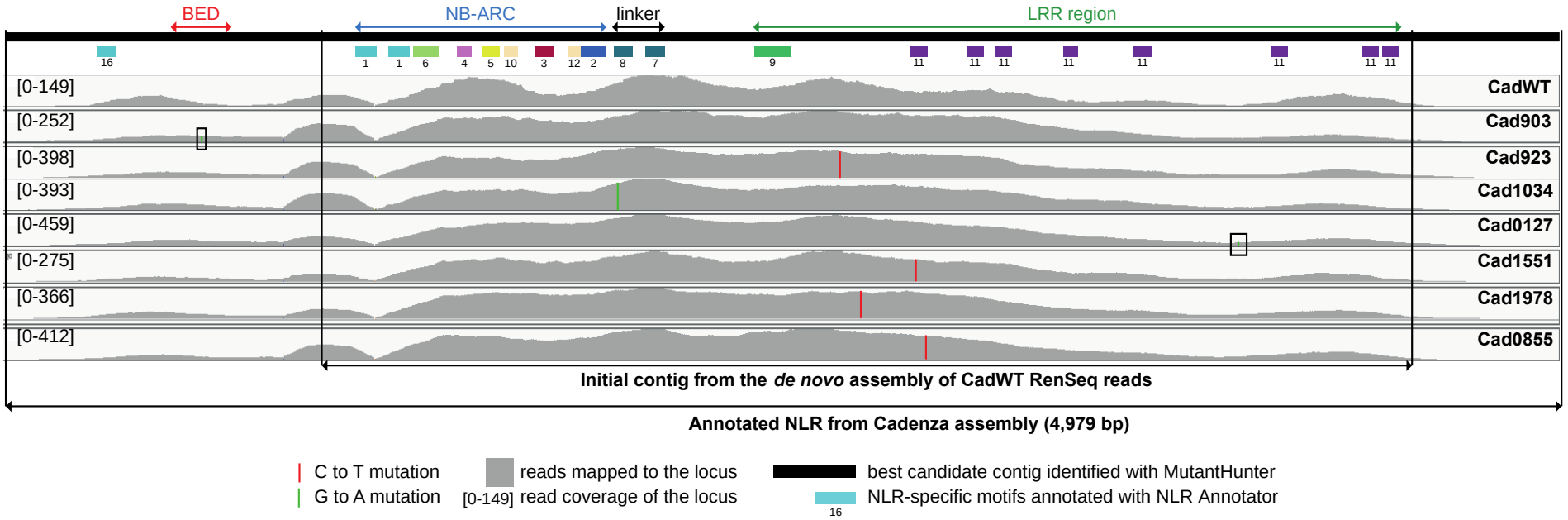


Figure 3-2. Identification of a candidate contig for *Yr7* using MutRenSeq.
View of RenSeq reads from the wild-type and EMS-derived Cadenza mutants mapped to the best *Yr7* candidate contig identified with MutantHunter. From top to bottom: vertical black lines represent the *Yr* loci, coloured rectangles depict the motifs identified by NLR-Annotator (each motif is specific to a conserved NLR domain), while read coverage (grey histograms) is indicated on the left, e.g. [0 - 149], and the line from which the reads are derived on the right, e.g. CadWT for Cadenza wild-type. Vertical bars represent the position of the SNPs identified between the reads and reference assembly – red shows C to T transitions and green G to A transitions. Black boxes highlight SNP for which the coverage was relatively low, but still higher than the 20x detection threshold. Vertical black lines illustrate the assembled candidate contigs and the one that was formerly de novo assembled from Cadenza RenSeq data, lacking the 5' region containing the Cad903 mutation

>*Yr7* candidate with Ns in Cadenza genome

ATGGAGCCGGCGGGAGACTCTTCCCTGGAGGCCGCGATTGCATGGCTGGTGCAGACCATCCTTGCAACGCTCCTCATGGACAAGATGGAGGCCTGGATTCAGCAAGTCGGGCTTGCCGACGACGTCGAGAGGCTCCAGTCTGAGGTCGAGAGAGTCGACACGGTGGTGGCTGCTGTGAAGGGGAGGGCAGCCGGGAACATGC
CTCTGTCCCGGTCTCTCGCTCGTGTCAAGGAGCTTCTCTATGACGCCGACGACGTGATCGACGAGCTAGACTACTACAGGCTCCAACACCAAGTCGAAGGAGGTAGTAAGCATAATCCCATTATATATCGAATCTATGTGTGCTACTCAATAGTTTGATCTTAATTTCTGGTCCATGTTTCTTTTCGGCACAGTTACAAGTG
ACGAGCCTGACGGTATGCGTGGAGCTGAAAGAGTGGATGAAATATCAAGGGGCCATGTCGATACACTGAATGTCAGTGTTGGCAAATTACGGTCCCCGGTATGGGAACACTTCACCATCACAGAAACAACTATCGACGGGAAGCGTTCAAAAGCCAAATGTAAGTACTGTGGAAATGATTTTAATTGCGAAACGAAGACAAA
CGGGACTTCATCTATGAAAAAACATTTGGAGAAGGAGCATTCCGTGACTTGCACGAATAAATCTGCAGTGCACCCCCCAAACACTTCAAGGTACCAGCAGGAATTTATACCTTGCTTCAACGAATTTGTTGTAATTGTTTATATACGTCTGCTTGAGAGCCCATTGTTGTTCTGAATTTCTTCTGATAACCAACCCACCATC
CTTTTCTTACTGCAGCACCGGCGATGCTACTTGTAATGTGAGGTCGGTTGAAGTTGGTAGTTCGTCCAACGGAAAAGAAAGAGAACAAATGAGGATCCNAAGGCTGAGTTATCCAATAGGATAATTAAAATTACTGAGAAGTTACAGTTACAGGACATCCAGGGGGCTTTGAGTAAAGTTCTCGAGCCATATGGATCCAGC
GCTACTTCAAGTTCAAATCATCACCGCTTGAGTACAGCATCAGATCAGCACCCAACAACATCAAGTCTTGTTCCAATGGAAGTTTATGGCAGAGTTGCAGAAAAGAATAAGATCAAAAAGTCAATAACTGAAAACCAATCTGGTGGTGTAAATGTTCTGCCTATTGTAGGCATTGCAGGTGTTGGAAAGACAACTCTTGCTC
AATTTGTGTATAATGATCCAGACNCAGAAAAGAATAAGATCAAAAAGTCAATAACTGAAAACCAATCTGGTGGTGTAAATGTTCTGCCTATTGTAGGCATTGCAGGTGTTGGAAAGACAACTCTTGCTCAATTTGTGTATAATGATCCAGACGTGAAAAGTCAATTTCACCACAGGATATGGGTTTGTGTGTCCTGCAAATT
TGATGAAGTGAAGCTCACAAAGGAGATGTTAGACTTTTTCCTCGAGAAAGGCATGAAGGAATTAACAACTTCGCGAAGCTTCAAGAGATCTTGAAAGAACATGTCGAGTACCAAGCAAAGAGTTTTCTGCTCATTTATAGATGATGTCTCGGACAGTATGGATTATCATAAATGGAACAAATTGTTGAACCCTTTGCTATCA
AGTCAAGCGAAGAATATAATTCTAGTCACGACCAGAAATTTGTCTGTTGCACAAAGGTTAAGCACACTTGAACCGATCAAGTTAGGTGCTTTAGAAAACGATGATATGTGGTTATTGCTCAAGTCATGTGCATTTGGTTTTGGGAACTATGAAGGTACGGAAAATCTAAGCACTATTGGAAGACAAATAGCAGAGAAGTTAA
AGGGCAATCCGTTAGCAGCAGTAACTGCAGGGGCACTGTTAAGAGATAATCTTAGCATTGATCATTGGAGTAACATTCTCAAGAATGAGAAGTGGAAATCGCTGGGACTCAGTGGGGGCATCATGCCTGCTTTGAAGCTTAGTTATGATGAGTTGACGTACCGTTTACAACAATGTTTCTCGTATTGCTCTATATTTCCTGA
CAAATATAGGTTTCTCGGGAAGGATTTGGTCTATATTTGGATTTCTCAGGGATTTGTGAATTGCACCCAAAATAAGAGATTGGAGGAGACGGGATGGGAATATCTGAATCAATTGGTAAATCTTGGATTCTTTCAACAAATTGAAGAACAACAAGAATTGGATGGGGAAGAAGAATTCTCTCTACGCCGTCAGATTTGGTAC
TCTATGTGGATCTCATGCATGATTTCGCAAGGATGATTTCAAGGACTGAATGTGCGACTATAGATGGTCTACAGTGCAATAAAATATTCCCAACTGTACAGCATTTGTCAATAGTAACCGGTTCTGCATACAACAAAGATCTGAAGGGGAACATTCCTCGTAATGAGAAGTTTGAAGAAAATATGAGAAATTCAGTTACAT
CAGTTACCAAATTGAGAACATTGGTTGTGCTTGGGAACTTTGACTCTTTCTTTGTACGGTTGTTCCAAGATATATTCCAGAAGGCACAAAATTTACGCCTGCTGCTAGTATCTCTAGCATCCACTTATCTGTCTCAAGTGCCTGCTGCATTCAATGATTTTAATTCCTTCCTGTGCAATTTGGCAAATCCTTTGCATCTTCG
TTACCTAAAACTTGAGTTGGATGGGATTGTGCCACAAGTTTTGAGTACGTTTTTTCATCTTCAAGTATTAGATGTTGGATCAAGCATGGATACTTCTCTACCCAATGGCTTGTTGCATAATCTTGTTAGCCTGCGACATCTTGTTGCACACAAGAGAGTCCATTCTTCCATTACTAGCATTGGTAACATGACATCTATCCAG
GAGCTACATGATTTTGAAGTTCGAATTTCTAGCGGCTTTGAGATAACACGACTCCAATCCATGAACGAGCTTGTTCAACTTGGGTTGTCTCAACTTGACAGTGTTAAAACCAGGGAGGACGCTTATGGGGCAGGACTAAGAAACAAGGAACACTTAGAAGAGCTTCATTTGTCCTGGAAGGATGCATATTCAGAGTATGAGT
ATGCCAGTGACACTGAATTTGAATCTTCTGCAAACATGGCAAGAGAAGTGATTGAGGGTCTTGAACCACACATGGATTTAAAACATCTACAAATATCTCAGTATAATGGTACCACTTCACCAGCTTGGCTTGCCAACAATATCTCAGTTACCTCATTGCAGACGCTTCATCTTGATGATTGTGGAGGATGGAGAATACTTCC
ATCTCTGGGAAGTCTTCCATTCCTTACAAAGGTGAAGTTGAGCAGCATGCTGGAAGTAATTGAAGTACTGATTCCTTCACTGGAGGAGCTAGTTCTAATTAAAATGCCGAAGTTAGTGAGATGCTCAAGCACTTCTGCCGAGGGTCTGAGCTCTAGCTTAAGGGTACTGCACATTGAGGATTGTGAAGCATTGAAGGAGTTT
GATCTGTTTGAGAACGATTATAATTCTGAAATCATTCAGGGATCATGGCTGCCTGGTCTTAGGAATTTGATTCTATATTGTTGCCCTCATTTGAAAGTGTTGAAGCCTCTTCCACCTTCAACTACCTTTTCTAAGGTACTCATCAGAGAAATTTCAAGATTTCCGTCTATGGAGGTATCATCTGGTGAGAAGTTACAAATTG
GGAATATTGATGTGTACATAGGCGATGATTTTGATGAGTCTTCTGATGAGTTGAGCATACTGGATGACAAAACTTTGGCGTTCCATAATCTTAGAAACCTGAAATCGATGGAGATATATGGTTGCAGAAATCTAAGGTCTTTTTCGTTCGAAGGTTTCAGTCATCTTGTCTCTTTAACAAGTTTGAAAATAGTAAGCTGTGA
ACAACTTTTCCCTTCAGATGTGACGGCAGAGTATACCCTTGAAGATGTGACAGCTGTGAACTGCAATGCCTTCCCATATCTTAAAAGCCTCAGTATCGACTCATGTGGAATAGCGGGGAAGTGGCTATCGCTGATGCTGCAGCATGCGCCAGGCCTAGAGGAATTGAGTTTAACAAGTTGCGCCCATATAACAAGAGTAGTG
TTACCGATGGAAGAGGAAGAAAACAATCTATTAACAACAGTACTGTCATCAGGAAATCAAGATGAGGCATTGACATGGTTAGTTCGTGACGGACTCTTGCACATTCCATCAAATCTCGTCTCCTCTCTCAAGAATATGAGTATTACTCAGTGCCCTCGCCTAAAGTTTAACTCAGGCAAGGACTGCTTCTCTGGATTTACCT
CGCTTGAGAAGCTTGAAATTTGGGGATCGTTGGTGGATGATGACGGAAGTGATGACCTGGAGAATGGAAGTTCTTTTGTGTTCGGAGAGGAGGATCAACCCCTGGGGGCGAACGGAAGATGGCTCCTCCCGACATCACTTCAGGAACTTCACATCGTGTCATTGTATTGCCAAGAAACGCTGCAAGTCTGCTTCCCTAGAGA
TATCACCAGCCTTAAAAAGTTAAGTGTACGTTCCGGCCAAGGTTTGCAATCTCTACAGCTGTACTCATGCACGGCACTGGAAGAATTGGCAATTTCCGGCTCTGGATCGGTCACCGTCACTGTACTAGAGGGCACGCAACCCGCTGGCAGCCTCGGGCGTTTGAATGTATCAGACTGTCCTGGCTTGCCATCACGTTTGGAC
AGCTTTCCAAGGTTGTGCCCTCGGCTGGAAAGGCTTGACATCAATGACCCATCTGTCCTTACCACGCCATTCTGCAAGCACCTCACCTCCCTGCAACGCCTAAAACTTGGCTTCTTGAAAGTGACGAGACTAACAGATGAGCAAGAACGAGCGCTTGTGCTCCTCAAGTCACTGAAAGAGCTCGAGATTTTTTATTGTACTC
ATCTCATAGATCTTCCTGCGGGGCTGCAGACCCTTCCTTCCCTCAAGAGTTTGAAGATAGAAGAGGGTCGAGGCATCTCAAGGCTGCCGGAAGCAGGCCTCCCACATTCGCTGGAAGAACTGGAAATCAAAATTTGCAGCAAGCTAGAAGATGAATGCAGGCGGCTAGCAACATGCGAAGGCAAGCTAAAAGTCAAAATTGA
TGGTCGATATGTGAATTAA

>Curated_Yr7

ATGGAGCCGGCGGGAGACTCTTCCCTGGAGGCCGCGATTGCATGGCTGGTGCAGACCATCCTTGCAACGCTCCTCATGGACAAGATGGAGGCCTGGATTCAGCAAGTCGGGCTTGCCGACGACGTCGAGAGGCTCCAGTCTGAGGTCGAGAGAGTCGACACGGTGGTGGCTGCTGTGAAGGGGAGGGCAGCCGGGAACATGC
CTCTGTCCCGGTCTCTCGCTCGTGTCAAGGAGCTTCTCTATGACGCCGACGACGTGATCGACGAGCTAGACTACTACAGGCTCCAACACCAAGTCGAAGGAGGTAGTAAGCATAATCCCATTATATATCGAATCTATGTGTGCTACTCAATAGTTTGATCTTAATTTCTGGTCCATGTTTCTTTTCGGCACAGTTACAAGTG
ACGAGCCTGACGGTATGCGTGGAGCTGAAAGAGTGGATGAAATATCAAGGGGCCATGTCGATACACTGAATGTCAGTGTTGGCAAATTACGGTCCCCGGTATGGGAACACTTCACCATCACAGAAACAACTATCGACGGGAAGCGTTCAAAAGCCAAATGTAAGTACTGTGGAAATGATTTTAATTGCGAAACGAAGACAAA
CGGGACTTCATCTATGAAAAAACATTTGGAGAAGGAGCATTCCGTGACTTGCACGAATAAATCTGCAGTGCACCCCCCAAACACTTCAAGGTACCAGCAGGAATTTATACCTTGCTTCAACGAATTTGTTGTAATTGTTTATATACGTCTGCTTGAGAGCCCATTGTTGTTCTGAATTTCTTCTGATAACCAACCCACCATC
CTTTTCTTACTGCAGCACCGGCGATGCTACTTGTAATGTGAGGTCGGTTGAAGTTGGTAGTTCGTCCAACGGAAAAGAAAGAGAACAAATGAGGATCCAACGCAGACCACCGCAGCTAACATACACGCCCAATGGGACAAGGCTGAGTTATCCAATAGGATAATTAAAATTACTGAGAAGTTACAGTTACAGGACATCCAG
GGGGCTTTGAGTAAAGTTCTCGAGCCATATGGATCCAGCGCTACTTCAAGTTCAAATCATCACCGCTTGAGTACAGCATCAGATCAGCACCCAACAACATCAAGTCTTGTTCCAATGGAAGTTTATGGCAGAGTTGCAGAAAAGAATAAGATCAAAAAGTCAATAACTGAAAACCAATCTGGTGGTGTAAATGTTCTGCCTA
ATTGTAGGCATTGCAGGTGTTGGAAAGACAACTCTTGCTCAATTTGTGTATAATGATCCAGACGTGAAAAGTCAATTTCACCACAGGATATGGGTTTGTGTGTCCTGCAAATTTGATGAAGTGAAGCTCACAAAGGAGATGTTAGACTTTTTCCTCGAGAAAGGCATGAAGGAATTAACAACTTCGCGAAGCTTCAAGAGAT
CTTGAAAGAACATGTCGAGTACCAAGCAAAGAGTTTTCTGCTCATTTTAGATGATGTCTCGGACAGTATGGATTATCATAAATGGAACAAATTGTTGAACCCTTTGCTATCAAGTCAAGCGAAGAATATAATTCTAGTCACGACCAGAAATTTGTCTGTTGCACAAAGGTTAAGCACACTTGAACCGATCAAGTTAGGTGCT
TTAGAAAACGATGATATGTGGTTATTGCTCAAGTCATGTGCATTTGGTTTTGGGAACTATGAAGGTACGGAAAATCTAAGCACTATTGGAAGACAAATAGCAGAGAAGTTAAAGGGCAATCCGTTAGCAGCAGTAACTGCAGGGGCACTGTTAAGAGATAATCTTAGCATTGATCATTGGAGTAACATTCTCAAGAATGAGA
AGTGGAAATCGCTGGGACTCAGTGGGGGCATCATGCCTGCTTTGAAGCTTAGTTATGATGAGTTGACGTACCGTTTACAACAATGTTTCTCGTATTGCTCTATATTTCCTGACAAATATAGGTTTCTCGGGAAGGATTTGGTCTATATTTGGATTTCTCAGGGATTTGTGAATTGCACCCAAAATAAGAGATTGGAGGAGAC
GGGATGGGAATATCTGAATCAATTGGTAAATCTTGGATTCTTTCAACAAATTGAAGAACAACAAGAATTGGATGGGGAAGAAGAATTCTCTCTACGCCGTCAGATTTGGTACTCTATGTGGATCTCATGCATGATTTCGCAAGGATGATTTCAAGGACTGAATGTGCGACTATAGATGGTCTACAGTGCAATAAAATATTC
CCAACTGTACAGCATTTGTCAATAGTAACCGGTTCTGCATACAACAAAGATCTGAAGGGGAACATTCCTCGTAATGAGAAGTTTGAAGAAAATATGAGAAATTCAGTTACATCAGTTACCAAATTGAGAACATTGGTTGTGCTTGGGAACTTTGACTCTTTCTTTGTACGGTTGTTCCAAGATATATTCCAGAAGGCACAAA
ATTTACGCCTGCTGCTAGTATCTCTAGCATCCACTTATCTGTCTCAAGTGCCTGCTGCATTCAATGATTTTAATTCCTTCCTGTGCAATTTGGCAAATCCTTTGCATCTTCGTTACCTAAAACTTGAGTTGGATGGGATTGTGCCACAAGTTTTGAGTACGTTTTTTCATCTTCAAGTATTAGATGTTGGATCAAGCATGGATACTTCTCTACCCAATGGCTTGTTGCATAATCTTGTTAGCCTGCGACATCTTGTTGCACACAAGAGAGTCCATTCTTCCATTACTAGCATTGGTAACATGACATCTATCCAG
GAGCTACATGATTTTGAAGTTCGAATTTCTAGCGGCTTTGAGATAACACGACTCCAATCCATGAACGAGCTTGTTCAACTTGGGTTGTCTCAACTTGACAGTGTTAAAACCAGGGAGGACGCTTATGGGGCAGGACTAAGAAACAAGGAACACTTAGAAGAGCTTCATTTGTCCTGGAAGGATGCATATTCAGAGTATGAGT
ATGCCAGTGACACTGAATTTGAATCTTCTGCAAACATGGCAAGAGAAGTGATTGAGGGTCTTGAACCACACATGGATTTAAAACATCTACAAATATCTCAGTATAATGGTACCACTTCACCAGCTTGGCTTGCCAACAATATCTCAGTTACCTCATTGCAGACGCTTCATCTTGATGATTGTGGAGGATGGAGAATACTTCC
ATCTCTGGGAAGTCTTCCATTCCTTACAAAGGTGAAGTTGAGCAGCATGCTGGAAGTAATTGAAGTACTGATTCCTTCACTGGAGGAGCTAGTTCTAATTAAAATGCCGAAGTTAGTGAGATGCTCAAGCACTTCTGCCGAGGGTCTGAGCTCTAGCTTAAGGGTACTGCACATTGAGGATTGTGAAGCATTGAAGGAGTTT
GATCTGTTTGAGAACGATTATAATTCTGAAATCATTCAGGGATCATGGCTGCCTGGTCTTAGGAATTTGATTCTATATTGTTGCCCTCATTTGAAAGTGTTGAAGCCTCTTCCACCTTCAACTACCTTTTCTAAGGTACTCATCAGAGAAATTTCAAGATTTCCGTCTATGGAGGTATCATCTGGTGAGAAGTTACAAATTG
GGAATATTGATGTGTACATAGGCGATGATTTTGATGAGTCTTCTGATGAGTTGAGCATACTGGATGACAAAACTTTGGCGTTCCATAATCTTAGAAACCTGAAATCGATGGAGATATATGGTTGCAGAAATCTAAGGTCTTTTTCGTTCGAAGGTTTCAGTCATCTTGTCTCTTTAACAAGTTTGAAAATAGTAAGCTGTGA
ACAACTTTTCCCTTCAGATGTGACGGCAGAGTATACCCTTGAAGATGTGACAGCTGTGAACTGCAATGCCTTCCCATATCTTAAAAGCCTCAGTATCGACTCATGTGGAATAGCGGGGAAGTGGCTATCGCTGATGCTGCAGCATGCGCCAGGCCTAGAGGAATTGAGTTTAACAAGTTGCGCCCATATAACAAGAGTAGTG
TTACCGATGGAAGAGGAAGAAAACAATCTATTAACAACAGTACTGTCATCAGGAAATCAAGATGAGGCATTGACATGGTTAGTTCGTGACGGACTCTTGCACATTCCATCAAATCTCGTCTCCTCTCTCAAGAATATGAGTATTACTCAGTGCCCTCGCCTAAAGTTTAACTCAGGCAAGGACTGCTTCTCTGGATTTACCT
CGCTTGAGAAGCTTGAAATTTGGGGATCGTTGGTGGATGATGACGGAAGTGATGACCTGGAGAATGGAAGTTCTTTTGTGTTCGGAGAGGAGGATCAACCCCTGGGGGCGAACGGAAGATGGCTCCTCCCGACATCACTTCAGGAACTTCACATCGTGTCATTGTATTGCCAAGAAACGCTGCAAGTCTGCTTCCCTAGAGA
TATCACCAGCCTTAAAAAGTTAAGTGTACGTTCCGGCCAAGGTTTGCAATCTCTACAGCTGTACTCATGCACGGCACTGGAAGAATTGGCAATTTCCGGCTCTGGATCGGTCACCGTCACTGTACTAGAGGGCACGCAACCCGCTGGCAGCCTCGGGCGTTTGAATGTATCAGACTGTCCTGGCTTGCCATCACGTTTGGAC
AGCTTTCCAAGGTTGTGCCCTCGGCTGGAAAGGCTTGACATCAATGACCCATCTGTCCTTACCACGCCATTCTGCAAGCACCTCACCTCCCTGCAACGCCTAAAACTTGGCTTCTTGAAAGTGACGAGACTAACAGATGAGCAAGAACGAGCGCTTGTGCTCCTCAAGTCACTGAAAGAGCTCGAGATTTTTTATTGTACTC
ATCTCATAGATCTTCCTGCGGGGCTGCAGACCCTTCCTTCCCTCAAGAGTTTGAAGATAGAAGAGGGTCGAGGCATCTCAAGGCTGCCGGAAGCAGGCCTCCCACATTCGCTGGAAGAACTGGAAATCAAAATTTGCAGCAAGCTAGAAGATGAATGCAGGCGGCTAGCAACATGCGAAGGCAAGCTAAAAGTCAAAATTGA
TGGTCGATATGTGAATTAA

Figure 3-3. Correction of *Yr7* candidate in Cadenza assembly based on RenSeq data and Sanger Sequencing.
The two Ns (showed in red with yellow highlight) in the sequence above correspond to one 39bp insertion and the deletion of 129 bp in the corrected sequence below

3.3.2.2. Yr5 candidate

We identified a candidate carrying a mutation in the seven investigated Lemhi-Yr5 mutant lines for *Yr5*. Two lines shared the same mutation (Lem387 and Lem241, Appendix 8-4, Figure 3-4). Unlike the candidate contig we identified for *Yr7* from a *de novo* assembly, this contig seemed to encompass a complete NLR.

Our collaborator Jianping Zhang carried out an independent MutRenSeq approach on four AvocetS-Yr5 susceptible mutants and identified a single contig carrying a mutation in all four tested lines (Appendix 8-4,Figure 3-6). This contig shared 100 % identity with the one identified in the Lemhi-Yr5 background.

We thus identified a total of eleven independent mutant lines from two different background carrying a mutation in the *Yr5* candidate contig.

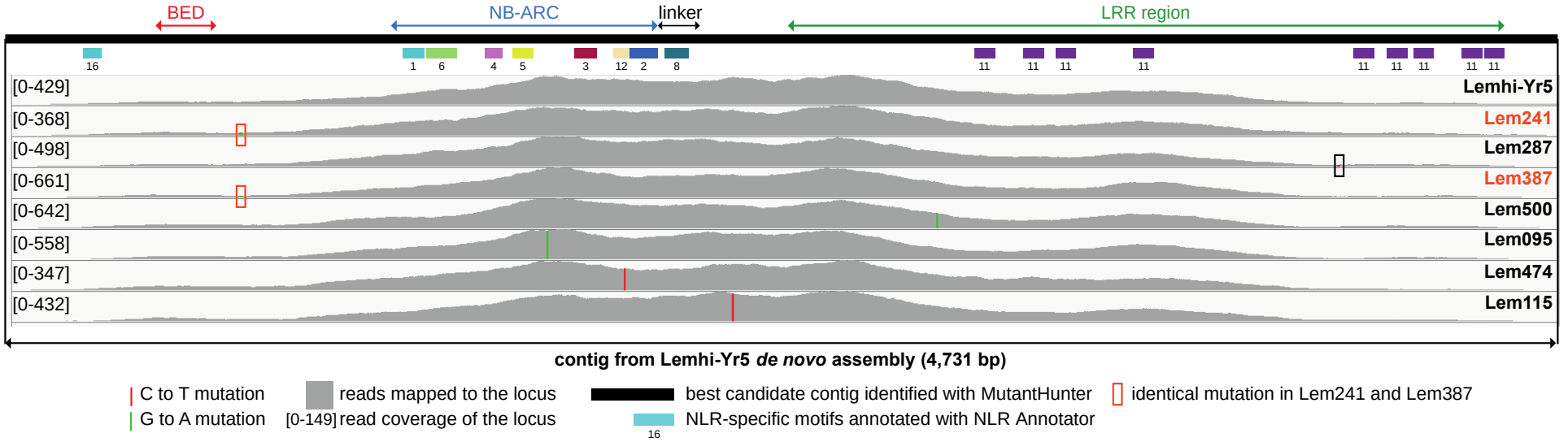


Figure 3-4. Identification of a candidate contig for *Yr5* using MutRenSeq.
View of RenSeq reads from the wild-type and EMS-derived Lemhi-Yr5 mutants mapped to the best *Yr5* candidate contig identified with MutantHunter. From top to bottom: vertical black lines represent the *Yr* loci, coloured rectangles depict the motifs identified by NLR-Annotator (each motif is specific to a conserved NLR domain), while read coverage (grey histograms) is indicated on the left, e.g. [0 - 149], and the line from which the reads are derived on the right. Vertical bars represent the position of the SNPs identified between the reads and reference assembly – red shows C to T transitions and green G to A transitions. Black boxes highlight SNP for which the coverage was relatively low, but still higher than the 20x detection threshold. Orange colour points out the two Lemhi-Yr5 mutant lines sharing the same mutation in the *Yr5* candidate contig (Lem241 and Lem387)

3.3.2.3. YrSP candidate

Our collaborator Jianping Zhang conducted a MutRenSeq approach similar to the one on AvocetS-Yr5 mutants to identify candidate contig for *YrSP* in four independent AvocetS-YrSP mutant lines. She identified one single contig carrying mutation in all four lines (Figure 3-5, Figure 3-6, Appendix 8-4)

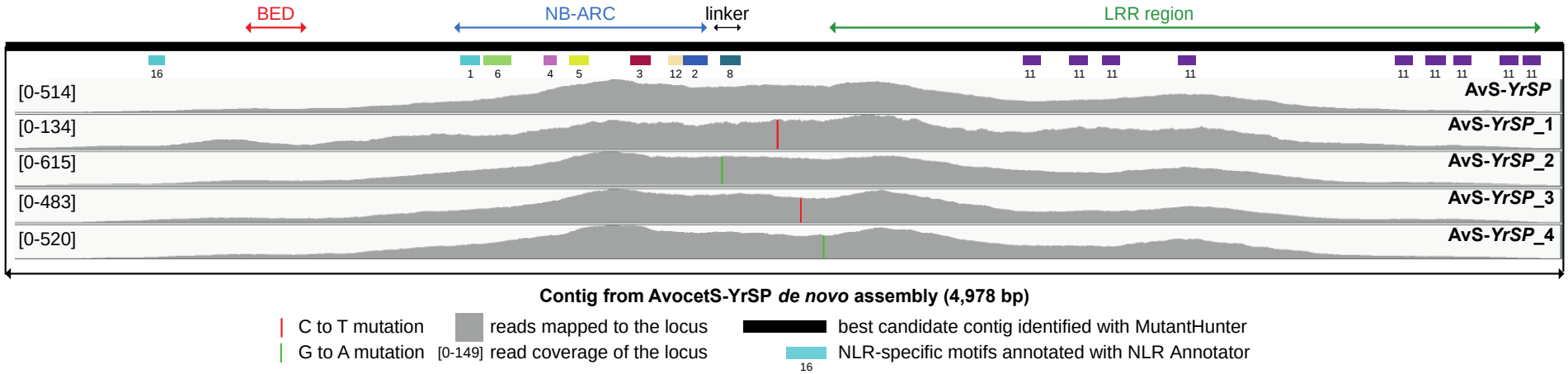


Figure 3-5. Identification of a candidate contig for *YrSP* using MutRenSeq.
View of RenSeq reads from the wild-type and EMS-derived AvocetS-YrSP mutants mapped to the best *YrSP* candidate contig identified with MutantHunter. From top to bottom: vertical black lines represent the *Yr* loci, coloured rectangles depict the motifs identified by NLR-Annotator (each motif is specific to a conserved NLR domain), while read coverage (grey histograms) is indicated on the left, e.g. [0 - 149], and the line from which the reads are derived on the right. Vertical bars represent the position of the SNPs identified between the reads and reference assembly – red shows C to T transitions and green G to A transitions. Black boxes highlight SNP for which the coverage was relatively low, but still higher than the 20x detection threshold. This analysis was carried out by Jianping Zhang (CSIRO).

3.3.2.4. Summary of the MutantHunter analysis to identify candidate contigs for Yr7, Yr5 and YrSP

In summary, we identified two strong candidate genes for both *Yr7* and *Yr5* (Figure 3-2 and Figure 3-4). These candidates were confirmed in an independent EMS-mutagenesis screen in AvocetS-Yr5 and AvocetS-Yr7 (Section 3.2.1, Appendix 8-4). In parallel our collaborators conducted a similar analysis to identify a candidate for *YrSP* (Figure 3-5).

Additionally, we conducted a BLAST analysis in the RenSeq *de novo* assemblies from AvocetS-Yr7, AvocetS-Yr5 and AvocetS-YrSP with each of the candidate contigs as query and each candidate was only present in the corresponding AvocetS-*Yr* introgression line and absent from AvocetS. This result provides further support to the hypothesis that the three candidate contigs encode *Yr7*, *Yr5* and *YrSP*. Interestingly, this analysis showed that the *Yr5* and *YrSP* contigs were almost identical with only two SNPs between the two. This was very surprising given that both *Yr5* and *YrSP* show very different resistance spectra to *Pst* (Chapter 2 - Introduction).

The next step was to determine whether the mutations that we identified in the candidate contigs would all lead to a variant in the predicted protein sequence.

3.3.3. Annotation of *Yr7*, *Yr5* and *YrSP* candidates

We used RNA-seq data already available for Cadenza (*Yr7*) and AvocetS-Yr5[197] to predict the coding region of both candidate genes and derived proteins (Figure 3-6). Details of the alignments are shown in Figure 3-12. These data allowed us to:

- (i) Determine the gene structure of the candidates
- (ii) Validate the natural variation between *Yr5* and *YrSP* (two SNPs identified from comparing the contigs in the above section) and predict its effect on the derived protein sequenced
- (iii) Predict the effect of the EMS mutations on the derived protein sequence

(i) Both genes share a similar gene structure with three exons and two introns. Exon 1 and 2 are approximately 300 bp each in both *Yr7* and *Yr5* and exon 3 is 4.1 kb in *Yr7* and 3.9 kb in *Yr5*. Introns 1 and 2 are approximately 120 bp in both candidates.

(ii) Given that *Yr5* and *YrSP* were almost identical we used the RNA-seq reads derived from AvocetS-Yr5 to defined *YrSP* candidate gene structure. Both exons 1 and 2 and introns 1 and 2 are identical to the *Yr5* candidate (Figure 3-6). We confirmed the two SNPs in *YrSP* when compared to the *Yr5* sequence in exon 3. The second SNP leads to a premature STOP codon in *YrSP*. Therefore exon 3 in *YrSP* is shorter (~ 2 kb) than in *Yr5*.

(iii) We first confirmed all the EMS mutations in the candidates by Sanger Sequencing. Then we used the gene structure defined in (i) to predict their effect on the protein sequence. Overall, all mutations were predicted to have an effect on the corresponding protein. We identified 18 amino-acid changes, four premature STOP codons and one mutation affecting the exon/intron junction (Figure 3-6).

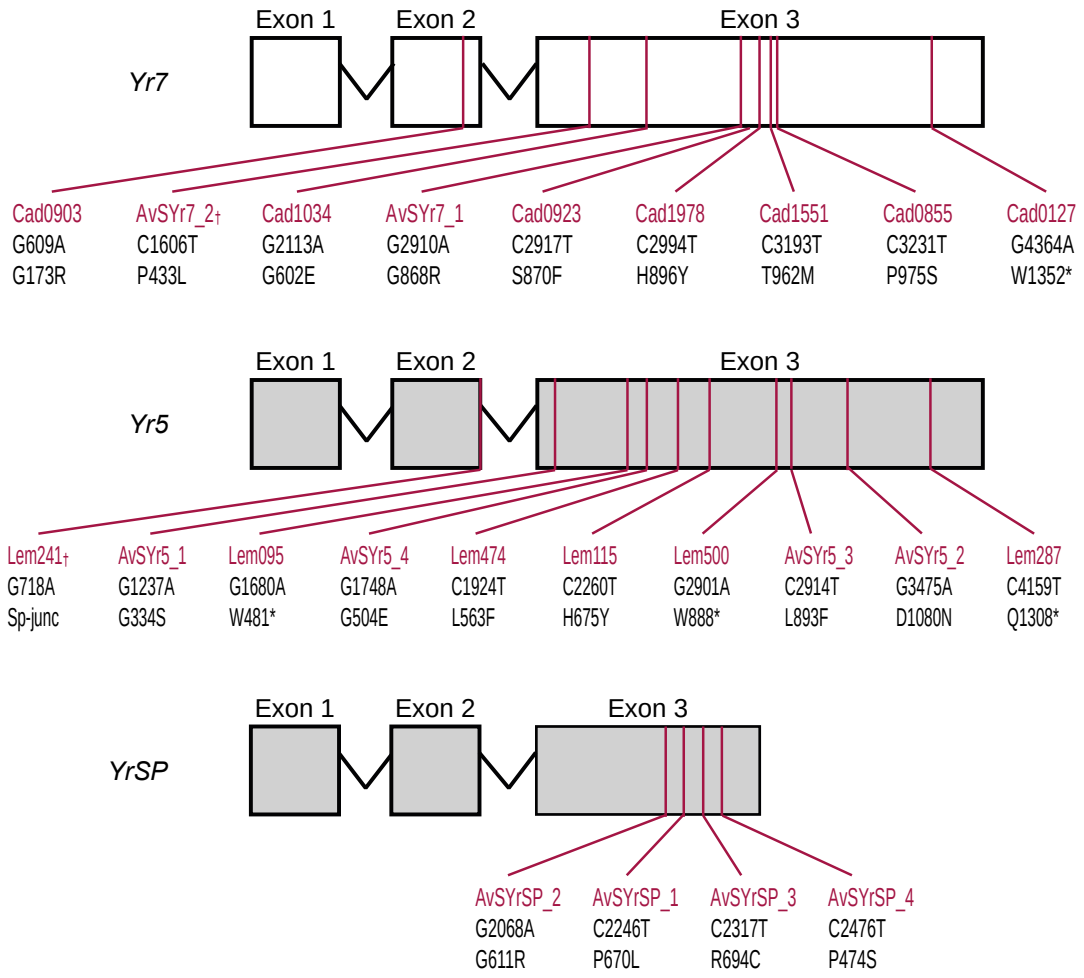


Figure 3-6. *Yr5* and *YrSP* are closely related sequences and distinct from *Yr7*. Mutations are shown in red with their predicted effects on the translated protein. Crosses show mutations shared by two independent mutant lines (Appendix 8-4). Background colour shows sequence similarity between the candidates: *Yr5* and *YrSP* share 99.8% identity, whereas *Yr7* and *Yr5* are 77.9% identical. Cad: Cadenza, Lem: Lemhi, AvSYr7: AvocteS-Yr7, AvSYr5: AvocetS-Yr5, AvSYrSP: AvocetS-YrSP.

Yr7 and *Yr5* share the same gene structure and a sequence similarity of 77.9 % across the coding sequence (Figure 3-6). Based on previously characterised resistance gene alleles in wheat, we hypothesized that *Yr7* and *Yr5* are two different genes based on sequence information. For instance, wheat *Pm3* alleles share a percentage identity greater than 97 %[209] and flax *L* alleles > 90 %[210]. Thus, based on sequence only, *Yr7* and *Yr5* are likely to be two different genes or distant paralogues. There were only two SNPs between *Yr5* and *YrSP*, one leading to an amino-acid change and the other is a single bp deletion in *YrSP* leading to a frameshift and premature termination codon in YrSP. Based on sequence comparison, *Yr5* and *YrSP* share 99.8 % identity, consistent with these two sequences being different alleles of a single gene.

3.3.4. Candidates are genetically linked to *Yr7*, *Yr5* and *YrSP* locus

To further confirm that these candidates are *Yr7*, *Yr5* and *YrSP*, we determined whether the candidates were genetically linked to the region where *Yr7*, *Yr5* and *YrSP* were initially mapped.

3.3.4.1. Investigating the link between the mutation in the Yr7 candidate in Cad127 and Cad1978 and the Yr7 loss of resistance phenotype

We carried out two different experiments to determine whether the mutations in the *Yr7* candidate were linked to the susceptible phenotype observed in the mutant lines:

- First, we investigated the segregation ratios between resistant and susceptible phenotype in F_2 progenies derived from two crosses between Cadenza wild-type and a *Yr7* loss of function mutant line (Cadenza wild-type x Cad127 and Cadenza wild-type x Cad1978).

- Then we performed bulked segregant analysis followed by exome capture and sequencing of the bulk in the F_2 population derived from Cadenza wild-type x Cad1978 cross.

Resistant:Susceptible phenotype segregation ratios in F_2 populations derived from Cadenza wild type x Cad127 and Cadenza wild type x Cad1978 (*Yr7*)

Here we will focus on two F_2 populations: Cadenza wild type x Cad127 and Cadenza wild type x Cad1978 (Table 3-3). We hypothesized that *Yr7*-mediated resistance is driven by a single-dominant gene. Hence, we expect a 3:1 (Resistant:Susceptible) segregation ratio in F_2 progenies derived from a cross between the wild-type parent and the homozygous loss of function mutant. We thus investigated F_2 progenies derived from the two crosses mentioned above and hypothesized that we should observe a 3:1 segregation ratio between resistant and susceptible individuals, respectively.

We screened 192 individuals derived from both Cadenza wild type x Cad127 and Cadenza wild type x Cad1978 crosses with PST 08/21 and determined the segregation ratio of resistant:susceptible phenotypes (Table 3-8). Both Cad127 and Cad1978 parents showed susceptible phenotypes (from 1 to 2^+), whereas the wild type Cadenza line showed a characteristic *Yr7* resistant phenotype (from 0nc to 1^-). No segregation ratio significantly close to 3:1 or 9:3:3:1 was observed in F_2 progenies (Table 3-8, data not shown for 9:3:3:1). A large number of plants did not show any symptoms, did not germinate or were too weak to score (82/192 and 39/192 for Cad127 and Cad1978. respectively). An inoculation issue might thus be the reason why the segregation ratios observed for Cad127 and Cad1978 were not consistent with the single gene 3:1 expected segregation. Indeed, we saw in Chapter 2 that the *Yr7* response is characterised by chlorotic and/or necrotic spots on the infected leaves and most of the individuals

phenotyped as resistant were actually not showing any symptom. Background mutations linked to the *Yr7* mutation could also affect segregation ratios and cause the observed distortion as plants with developmental issues could not be inoculated/scored.

Table 3-8. Segregation ratios observed in Cadenza wild-type x Cad127 and Cadenza wild-type x Cad1978 and comparison with the expected number in a 3:1 ratio scenario.

	Phenotypes		**Observed**			**Expected in 3:1**		**CHITEST**
Mutant Line	Wild type	Mutant parent	#Wild type	#Susceptible	Total	#Wild type	#Susceptible	*P value*
Cad127	0; to 1-	2	95	15	110	82.5	27.5	< 0.001
Cad1978	0; to 1-	1to2+	133	20	153	114.7	38.3	< 0.001

Given that the segregation analysis was not conclusive, we pursued the genetic mapping of the candidate *Yr7* mutation in Cad127 to determine whether it was genetically linked to the *Yr7* locus.

Exome-capture and sequencing on Cadenza wild-type x Cad1978 F_2 bulks (*Yr7*)

We conducted a bulked segregant approach for *Yr7* in the same F_2 population we investigated the segregation ratio between the resistant and susceptible phenotypes (Cadenza wild-type x Cad1978). We hypothesized that even though we did not observe the expected segregation ratio for a single dominant gene, we could still exploit the information from the susceptible (IT score of 2 at least) and resistant (IT score similar to 0nc) individuals to determine whether mutations that are enriched in the susceptible progenies were linked to chromosome arm 2BL (where *Yr7* was mapped).

We screened 153 individuals derived from Cad1978 x Cadenza wild-type cross with PST 08/21 and determined the segregation ratio of resistant:susceptible phenotypes (Table 3-8). We assembled bulks with equal amount of leaf tissue from the 20 susceptible individuals (presence of *Pst* pustules) and 20 resistant individuals showing a phenotype comprised between 0; and 0nc. We extracted DNA from these bulks and submitted them for exome capture and sequencing at the Earlham Institute.

We described in Section 3.2.6 the strategy to identify potential EMS-induced SNPs in the sequencing data from the bulks (Cadenza) when aligned to RefSeqv1.0 (Chinese Spring) and how we calculated the allele frequencies for wild-type and mutant alleles. We mentioned in Chapter 1 that a SNP linked to the susceptible phenotype would be enriched in the susceptible bulk, with an associated allele frequency close to 1, and depleted in the wild-type bulk with an allele frequency ranging between 0 and 0.25 as there might be heterozygous individuals in the wild-type bulk in this case.

We calculated the rolling average (seven SNP window size) of the allele frequencies for both bulks across the genome (Figure 3-7). We identified a difference in the allele

frequencies between the susceptible and wild-type bulks on chromosome 2B part1 and part2 (chr2B_part2) windows. However, the difference seems larger on the chr2B_part2 window where the susceptible bulk approached an allele frequency of 0.75 (three main peaks) whereas the resistant bulk had a lower mutant frequency of 0.25 (Figure 3-7). We added the two most distal markers we found in the literature that were linked to *Yr7* on the figure to determine whether the region where we observe the distortion in the allele frequencies coincide with the *Yr7* locus (WMS120 and WMS526, Figure 3-7). The locus partly overlaps with the region where we observed a distortion of the allele frequencies. The region of interest is wide given that the F_2 population screened included few individuals, therefore, the likelihood of obtaining recombination in the *Yr* interval is relatively low. This experiment thus provides evidence that the Cad1978 mutation located in the *Yr7* candidate is in the *Yr7* interval, although it does not provide direct evidence of the *Yr7* candidate being the actual *Yr7*.

To further confirm the genetic link between the *Yr7* candidate and the *Yr7* mapping interval on chromosome arm 2BL, we carried out a traditional mapping approach in Cadenza wild-type x Cad127 cross to determine whether markers linked to *Yr7* were also linked to the *Yr7* candidate (Figure 3-8). Additionally, we designed markers targeting the *Yr5* and *YrSP* candidates and carried out a similar mapping approach in AvocetS x AvocetS-Yr5 and AvocetS x AvocetS-YrSP F_2 progenies to investigate whether the *Yr5* and *YrSP* candidates were linked to markers linked to the *Yr5* and *YrSP* locus

.

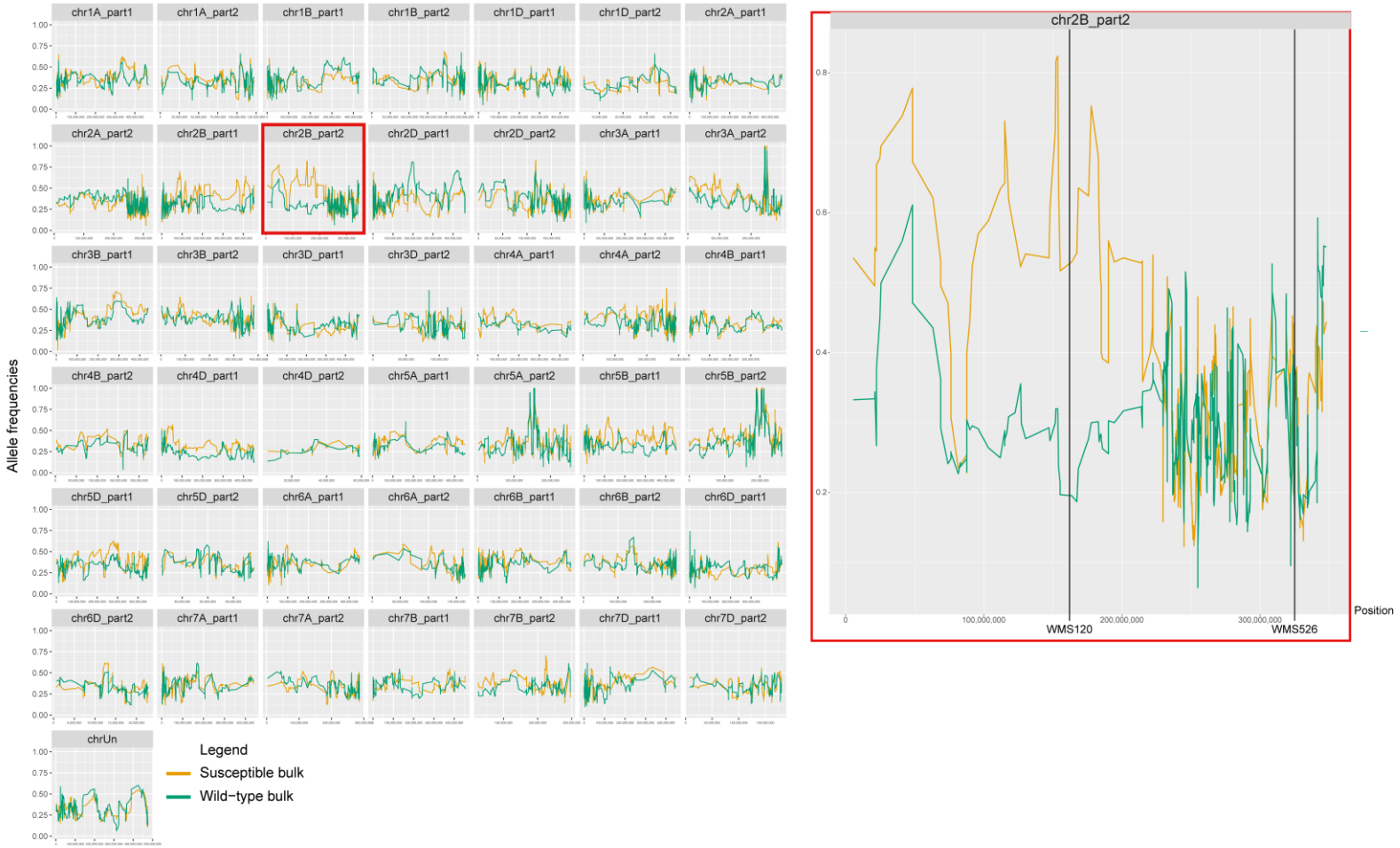


Figure 3-7. Distortion of allele frequencies on chromosome 2B (part 2) from Chinese Spring between susceptible and wild-type bulks (Cad1978)
Left: Allele frequencies (rolling average of 7 SNP positions) in susceptible (orange) and wild-type (green) bulks from Cadenza wild-type x Cad1978 F_2 population (153 individuals, 20 individuals per bulk) against RefSeqv1.0 (Chinese Spring) chromosome parts. Right: close-up on chromosome 2B part 2 overlapping with the *Yr* locus defined on (Figure 3-8)

3.3.4.2. *Traditional genetic mapping*

Yr7, *Yr5* and *YrSP* were initially mapped to chromosome arm 2BL[55,180]. We used previously published markers linked to *Yr7*, *Yr5*, and *YrSP* (WMS526 (*Yr7*), WMS501 (*Yr5*) and WMC175, WMC332 (*YrSP*)[180,203,204]) in addition to closely linked markers WMS120, WMS191, and WMC360 (GrainGenes database https://wheat.pw.usda.gov/GG3/) to define their physical mapping interval on the Chinese Spring assembly RefSeqv1.0[112] (Figure 3-8). Section 3.2.5 describes how the different KASP markers were designed and section 3.2.1 lists the F_2 populations we investigated for the different genes (Appendix 8-7, Table 3-3).

We found that the three candidate contigs were genetically linked to the intervals described above (Figure 3-8). More specifically, the *YrSP* candidate was fully linked to the M1 marker and tightly linked to the M2/M3 markers (0.9 cM, purple in Figure 3-8). *Yr5* was linked and flanked by both WMC175 and M3 (2.9 and 2.1 cM, respectively, red in Figure 3-8). *Yr7* was linked and flanked by both M1 and M2 markers (2.5 and cM, respectively, blue on Figure 3-8).

Interestingly, when projected onto the RefSeqv1.0 assembly, all three *Yr7*, *Yr5* and *YrSP* physical intervals partly overlap. This is consistent with previous work showing that these genes are fully linked and/or allelic. Indeed, no recombinant was previously found between *Yr7* and *Yr5* among 143 F_3 progenies[154], between *YrSP* and *Yr7* (208 F_3 lines) and between *YrSP* and *Yr5* (256 F_3 lines)[153]. Interestingly, the sequences with highest similarity in the Chinese Spring wheat genome sequence (RefSeq v1.0) all lie within this common physical interval (19 Mb wide, Figure 3-8).

In summary, these results provide genetic evidence that the three candidate genes are *Yr7*, *Yr5* and *YrSP*.

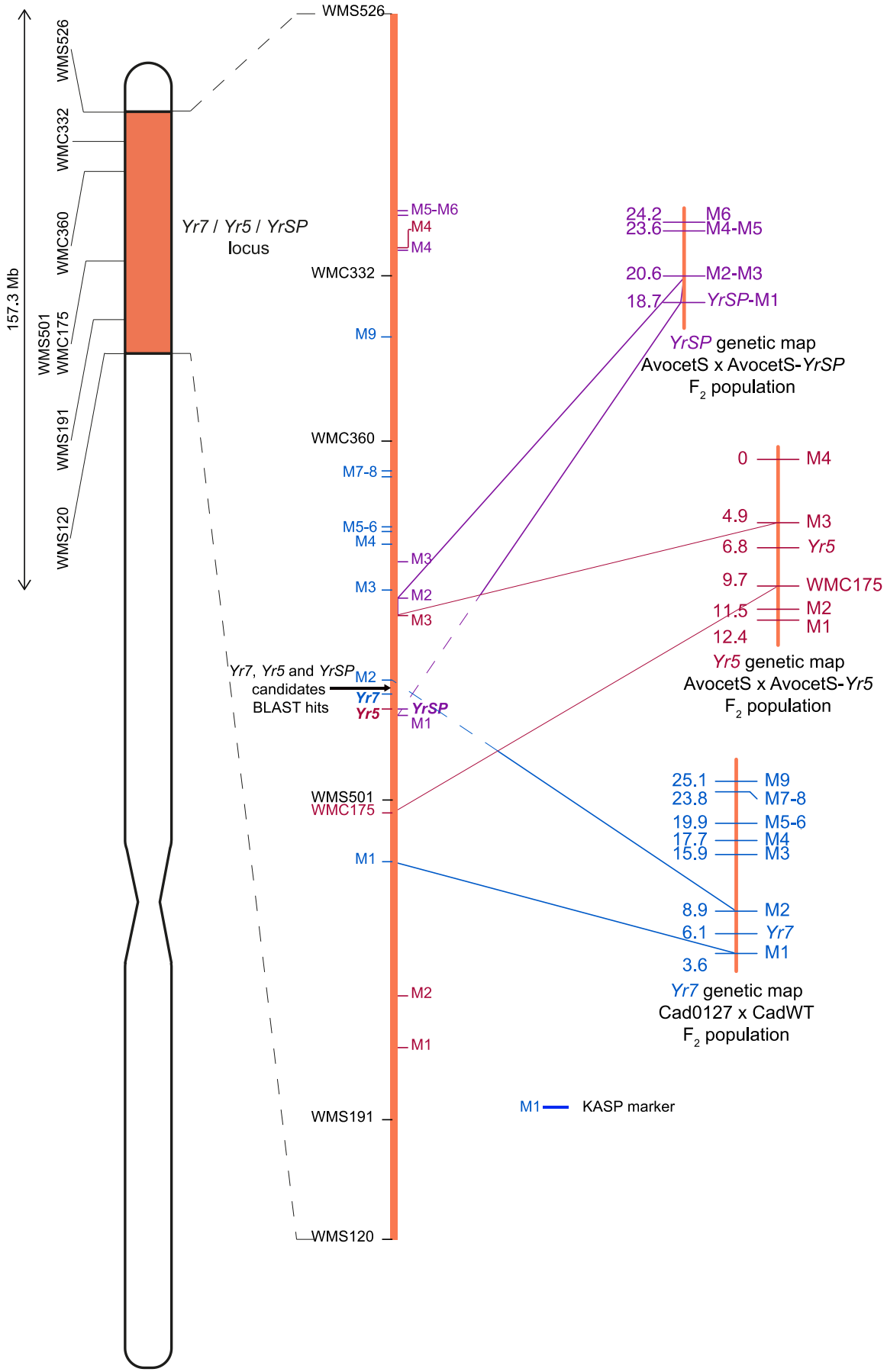

157.3 Mb
WMS526
WMC332
WMC360
WMS501
WMC175
WMS191
WMS120
Yr7 / Yr5 / YrSP
locus
WMS526
M5-M6
M4
M4
WMC332
M9
WMC360
M7-8
M5-6
M4
M3
M3
M2
M3
Yr7, Yr5 and YrSP
candidates
BLAST hits
M2
Yr7
Yr5
YrSP
M1
WMS501
WMC175
M1
M2
M1
WMS191
WMS120
24.2 M6
23.6 M4-M5
20.6 M2-M3
18.7 YrSP-M1
YrSP genetic map
AvocetS x AvocetS-YrSP
F2 population
0 M4
4.9 M3
6.8 Yr5
9.7 WMC175
11.5 M2
12.4 M1
Yr5 genetic map
AvocetS x AvocetS-Yr5
F2 population
25.1 M9
23.8 M7-8
19.9 M5-6
17.7 M4
15.9 M3
8.9 M2
6.1 Yr7
M1
3.6
Yr7 genetic map
Cad0127 x CadWT
F2 population
M1 KASP marker

Figure 3-8. Candidate contigs identified by MutRenSeq are genetically linked to the *Yr* mapping interval.
Schematic representation of chromosome 2B from Chinese Spring (RefSeq v1.0) with the positions of published markers linked to the *Yr* loci and surrounding closely linked markers that were used to define their physical position (orange rectangle). The chromosome is depicted as a close-up of the physical locus indicating the positions of KASP markers that were used for genetic mapping (horizontal bars, Appendix 8-7). Blue colour refers to *Yr7*, red to *Yr5*, and purple to *YrSP*. The black arrow points to the NLR cluster containing the best BLAST hits for *Yr7* and *Yr5*/*YrSP* on RefSeq v1.0. Coloured lines link the physical map to the corresponding genetic map for each targeted gene. Genetic distances are expressed in centiMorgans (cM).

Combining the segregation ratios analysis (Table 3-8) with the traditional mapping of the three candidate genes for *Yr7*, *Yr5* and *YrSP* (Figure 3-8) and the bulked segregant analysis in Cad1978 (Figure 3-7), we could demonstrate genetic linkage between our three candidate and the genetic interval of *Yr7*, *Yr5* and *YrSP*. In addition, all the mutations we identified in the candidates were predicted to have a deleterious effect on the predicted protein (Figure 3-6). Furthermore, each candidate was found only in the corresponding AvocetS-Yr line and not in the AvocetS recurrent parent. Based on these evidences which all support the three candidate contigs, we proceeded to further validate them. We will now refer to the candidate sequences as *Yr7*, *Yr5* and *YrSP*. Additionally, given that *Yr5* and *YrSP* have very similar sequences, we will now refer to them as *Yr5*/*YrSP*.

3.3.5. *Yr7*, *Yr5* and *YrSP* encode BED-NLRs

We predicted the proteins encoded by the *Yr7*, *Yr5*, and *YrSP* sequences (Figure 3-9). Interestingly, they encode non-canonical NLR proteins: they contain a zinc-finger BED domain at the N-terminus, followed by the canonical NB-ARC domain (Figure 3-9). Only the Yr7 and Yr5 proteins encode multiple LRR motifs at the C-terminus with YrSP having lost most of the LRR region due the premature termination codon in exon 3. *YrSP* still confers functional resistance to *Pst*, although with a recognition specificity different from *Yr5* (Chapter 2); all isolates virulent to *YrSP* are avirulent to *Yr5*, whereas the two isolates virulent to *Yr5* are avirulent to *YrSP*[157].

Yr7 and Yr5/YrSP are highly conserved in the N-terminus regions up to the BED domain (~ 95 % identity over 185 amino-acids). The BED domains itself is 51 amino-acids long and there is only one amino-acid change between Yr7 and Yr5 (Figure 3-9). This high degree of conservation is quickly eroded downstream of the BED domain with ~ 70 % identity from the end of the BED domain to the end of the NB-ARC. There is even more variation in the LRR region between Yr7 and Yr5, with the percentage of conservation varying from 0 to 85 %.

The BED domain is required for *Yr7*-mediated resistance, as a single amino acid change in mutant line Cad903 leads to a susceptible reaction (Figure 3-9). Given the presence of this non-canonical domain at the N-terminus of Yr7, Yr5 and YrSP, we hypothesized that it could be an integrated domain as described in the integrated decoy model[211]. However, recognition specificity is not solely governed by the BED domain, as *Yr5* and *YrSP* have identical BED domain sequences, yet confer resistance to different *Pst* isolates.

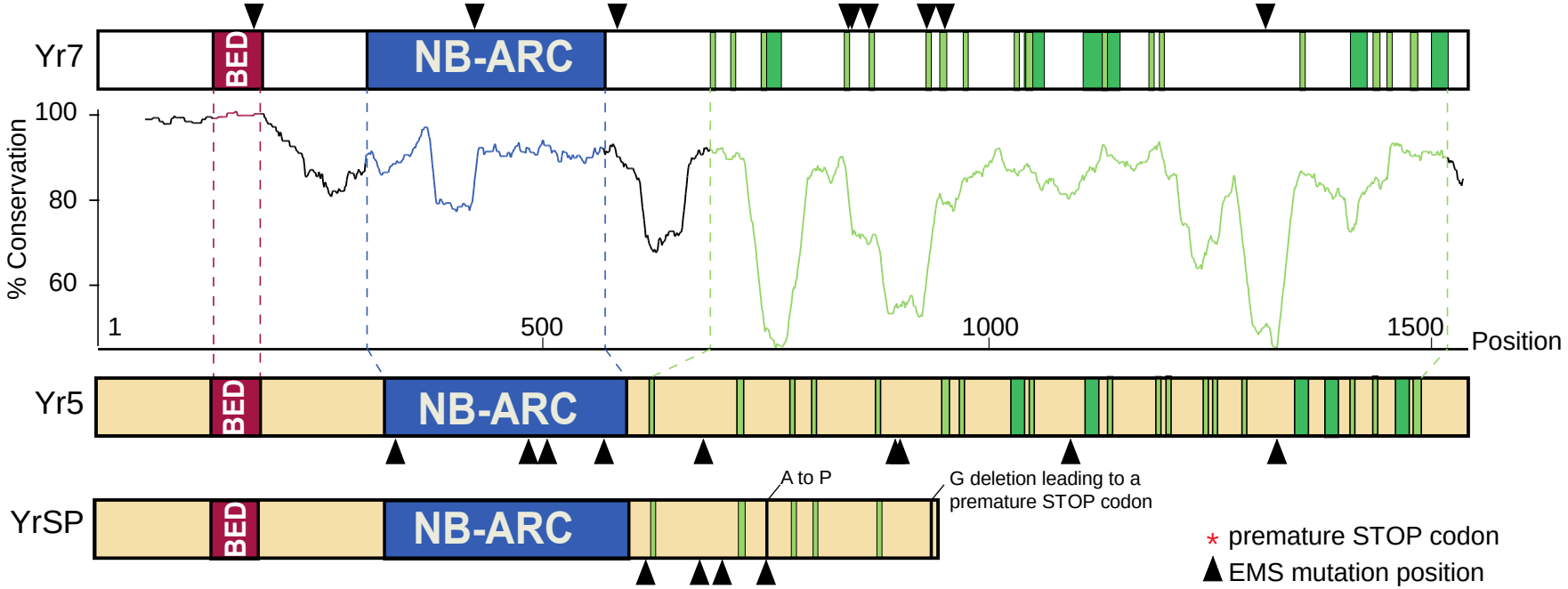


Figure 3-9. Schematic representation of the Yr7, Yr5, and YrSP protein domain organisation.
BED domains are highlighted in red, NB-ARC domains are in blue, LRR motifs from NLR-Annotator are in dark green, and manually annotated LRR motifs (xxLxLxx) are in light green. Black triangles represent the EMS-induced mutations within the protein sequence. The plot shows the degree of amino acid conservation (50 amino acid rolling average) between Yr7 and Yr5 proteins, based on the conservation diagram produced by Jalview[212] (2.10.1) from the protein alignment. Regions that correspond to the conserved domains have matching colours. The amino acid changes between Yr5 and YrSP are annotated in black on the YrSP protein

3.3.6. *Yr7, Yr5* and *YrSP* do not encode Coiled-Coil domains

3.3.6.1. Comparison of N-terminus sequence of Yr7 and Yr5/YrSP with characterised wheat CC-NLRs

Unlike previously cloned resistance genes in grasses (e.g. *Mla10*[213], *Sr33*[172], *Pm3*[214]), neither *Yr7* nor *Yr5*/YrSP encode Coiled Coil domains at the N-terminus (Figure 3-10). We compared Yr7 and Yr5/YrSP profiles with the COILS programs[199] to those obtained with already characterised CC-NLR encoding genes *Sr33, Mla10, Pm3* and *RPS5* (Figure 3-10). The 14 amino acid sliding window is the least accurate according to the COILs user manual, consistent with the additional peaks observed in Sr33, Mla10 and Pm3 that were not annotated as CC domains in the corresponding publications[172,209,215]. Thus, the peak at position 1,200 in Yr5 is unlikely to represent a CC domain.

To test the hypothesis that a putative CC domain was disrupted by the integration of the BED domain, we manually removed the BED domain peptide sequence and ran the modified protein sequence of Yr7 and Yr5 through the COILs program. There was no difference in the prediction between the two Yr proteins with or without their BED domain (Figure 3-10). We performed a BLASTP[216] search with the N-terminus region of the Yr5 and Yr7 proteins (from Met to the first amino-acid encoding the NB-ARC) with or without the BED domain and the best hits were proteins predicted to encode BED-NLRs from *Aegilops tauschii*, *Triticum urartu* and *Oryza sativa* (data not shown). Based on the COILS prediction and the BLAST search, we concluded that *Yr7* and *Yr5*/*YrSP* do not encode CC domains.

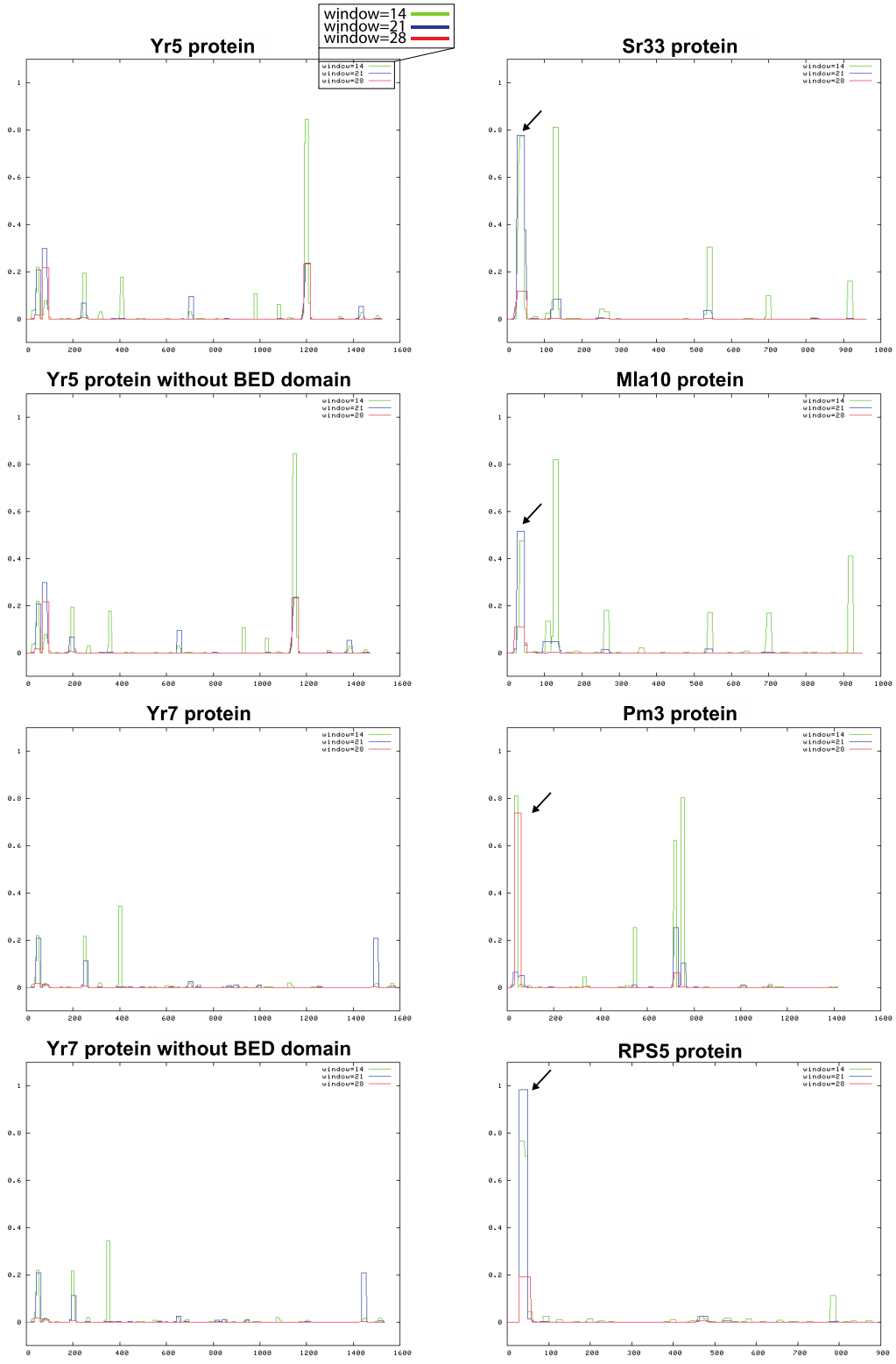


Figure 3-10. Yr7, Yr5 and YrSP proteins do not encode a Coiled-Coil domain in the N-terminus.
Graphical outputs from the COILS prediction program in three sliding windows (14, 21, and 28 amino acid, shown in green, blue, and red, respectively) for Yr5 and Yr7 with or without the BED domain (left) and characterized canonical NLRs: Mla10[213], Sr33[172], Pm3[214] and RPS5[217]. The X axis shows the amino acid positions and the Y axis the probability of a coiled coil domain formation. Black arrows point CC domain location.

3.3.6.2. Structure prediction of the Yr7 N-terminus and comparison with resolved structures of known CC-NLRs

To further investigate whether Yr7, Yr5 and YrSP possess a Coiled coil domain, we used Phyre2[200] to predict the protein structure based on sequence homology with deposited proteins and 3D model prediction. We will only show the results for Yr7 here, but identical conclusions were drawn from the Yr5/YrSP analysis. We selected the amino-acid sequence from the starting methionine to the start of the BED domain (135 amino-acids in *Yr7*), as it is the location of the CC domain in already cloned CC-NLRs (Figure 3-10). Surprisingly, first the program predicted a transmembrane helix between positions 8 and 26 and this was confirmed when using TMPred[218]. Secondly, despite not finding evidence for CC domain with the COILs program, Phyre2 found a high probability for Yr7 sequence to be homologous to both the N termini of RPP13-like protein 14 (CNL from *Arabidopsis thaliana*) and the CC domain from Mla10 (99.8 and 99.7 % probability, respectively, Figure 3-11).

However, when superimposing the structures, we could observe that only 59 atoms (~ 1,000 atoms in the total length of the sequence) were matched between Yr7 N-terminus and Mla10 and that the root mean square distance (RMSD) between the set of aligned atoms was 2.19 Å. This means that i) very few atoms actually matched between the two structures and that ii) the two structure do not perfectly overlap. The aligned region of RPP13-like protein 4 displayed a more similar fold to the Yr7 sequence (Figure 3-11). However, similar number of matching atoms and RMSD were found (75 atoms, 2.55 Å). Hence the high probability observed for Yr7 N-terminus to be homologous to these sequences could be due to a low number of available CC domains from NLRs in the database.

It is thus still unclear whether Yr7, Yr5 and YrSP truly encode a CC domain based on these observations. Similarly to what we performed above, we asked the question whether the integration of the BED domain could have disrupted a former CC domain. To answer this, we submitted the Yr7 amino-acid sequence from start codon to beginning of NB-ARC with BED domain deleted. There was no evidence of a complete CC domain in Yr7 and the aligned Yr7 N-terminus and Mla10/RPP13-4 sequences and structure were exactly the same as shown on Figure 3-11.

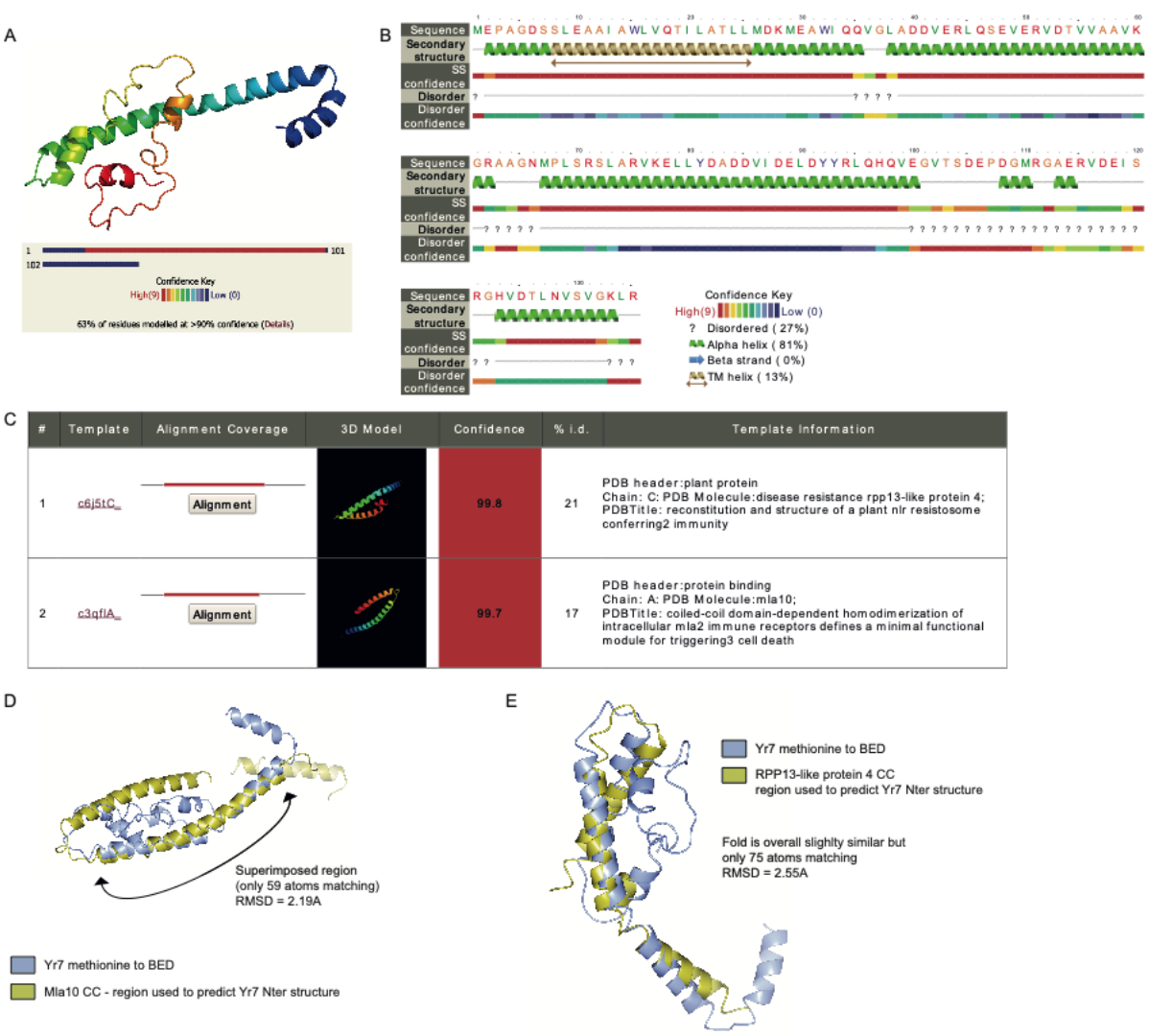


Figure 3-11. Structure prediction of Yr7 N-terminus with Phyre2.

A: Predicted structure of Yr7 from starting methionine to the beginning of the BED domain. Only 60 % of the sequence was predicted with high confidence based on two structures in the database, Mla10 CC and RPP13-like protein 4.

B: Predicted secondary structure showing the alpha helixes (green) and the predicted transmembrane region (light brown). The bar below shows the confidence of the prediction with red being close to 100% and blue to 0 %. The disorder track shows the regions that do not display any particular secondary structure and the track below this shows how likely the disorder is to be real. The confidence key is the same as for secondary structure prediction.

C: Table showing the two highest hits in the database to Yr7 N-terminus: Mla10 and RPP13-like protein 4. two sequences were also the ones used to predict the structure. The alignment column shows that only part of the Yr7 N-terminus sequence could be aligned with a percentage identity ranging from 17 to 21 %.

D: Screenshot of the superimposition of Mla10 and Yr7 N-terminus aligned regions with CCP4mg[219]. Only 59 atoms matched with a RMSD of 2.19 Å.

E: Screenshot of the superimposition of RPP13-like protein 4 and Yr7 N-terminus aligned regions with CCP4mg. Only 75 atoms matched with a RMSD of 2.55 Å.

3.3.7. *Yr7* and *Yr5*/*YrSP* variants are present in sequenced wheat cultivars

We aimed to design diagnostic markers for *Yr7*, *Yr5* and *YrSP* to use for marker-assisted selection in breeding programs. We cannot use the markers shown in Appendix 8-7 that we used for the genetic mapping given that they are derived from the EMS mutagenesis and are thus private to the mutant lines. Indeed, for *Yr7* we used the mutation information obtained from the sequenced TILLING line Cad127 as markers to be able to differentiate between Cadenza wild-type parent from Cad127 mutant parent (more details in Section 3.2.5).

For *Yr5* and *YrSP*, as we used F_2 populations derived from AvocetS-Yr NIL x AvocetS, we designed markers able to differentiate between the targeted candidate and its closest homolog in AvocetS. Such markers are thus only suitable in these specific bi-parental populations and more sequence information is required to develop diagnostic markers that could discriminate between the causal gene and its closest alleles/homologs in other wheat varieties worldwide. We thus examined the variation in *Yr7*, *Yr5*, and *YrSP* across eight sequenced tetraploid and hexaploid wheat genomes (Appendix 8-6).

For *Yr7*, we identified this sequence only in Cadenza and Paragon (Table 3-9). The Paragon assembly had exactly the same 'Ns' in *Yr7* as in the Cadenza assembly and we corrected the Paragon sequence with Sanger Sequencing in a similar manner as shown in Figure 3-3. Both cultivars are derived from the original source of *Yr7*, tetraploid durum wheat (*T. turgidum* ssp. *durum*) cultivar Iumillo and its hexaploid derivative Thatcher (Figure 3-15). None of the three sequenced tetraploid accessions (Svevo, Kronos, Zavitan) carry *Yr7* (Table 3-9).

Table 3-9. *In silico* allele mining for *Yr7* and *Yr5*/*YrSP* in available genome assemblies for wheat at the time of the study.
Table presents the percentage identity (% ID) of the identified variants and matching colours illustrate identical haplotypes. Investigated genome assemblies are shown in Appendix 8-6.

Cultivar	%ID to Yr5 protein	%ID to Yr7 protein
Cadenza	98.2	100
Paragon	98.2	99.8*
Claire	99.3	n.h
Robigus	98.2	n.h
Kronos	93.6	n.h
Svevo	93.6	n.h
Zavitan	n.h	n.h

* due to the presence of the Ns in the Paragon sequence
n.h means 'no hits' sharing more than 90 % identity

For *Yr5/YrSP*, we identified three additional haplotypes in the sequenced hexaploid wheat cultivars and could confirm the expression and gene structure of two of them with available RNA-Seq data (Figure 3-12 and Figure 3-13). Cultivar Claire encodes a complete NLR with nine amino-acid changes, including four polymorphisms in the C-terminus compared to Yr5. Cultivars Robigus, Paragon, and Cadenza also encode a full length NLR that shares common polymorphisms with Claire, in addition to 17 amino acid substitutions across the BED and NB-ARC domains. This haplotype was confirmed in RNA-Seq data from Cadenza (Figure 3-12). The presence of the *Yr5*/*YrSP* haplotype in Cadenza, which also carries *Yr7*, further supports the non-allelic relationship of these genes. The C-terminus polymorphisms between Yr5 and the other cultivars is due to a 774 bp insertion in *Yr5*, close to the 3' end, which carries an alternate termination codon (Figure 3-16).

Tetraploid cultivars Kronos and Svevo encode a fifth Yr5/YrSP haplotype with a truncation in the LRR region distinct from YrSP, in addition to 31 amino acid

substitutions across the C-terminus (Figure 3-13). This truncated tetraploid haplotype is reminiscent of YrSP and is expressed in Kronos (Figure 3-12). However, none of these cultivars (Claire, Robigus, Paragon, Cadenza, Svevo, and Kronos) exhibit a *Yr5*/*YrSP* resistance response, suggesting that these amino acid changes and truncations may alter recognition specificity or protein function. Additional testing of these haplotypes will provide insight into whether they represent a functional allelic series.

With this sequence information, we could design specific primers to differentiate between *Yr5* and its alternate alleles, *YrSP* and its alternate alleles, and *Yr7* (Section 3.3.8).

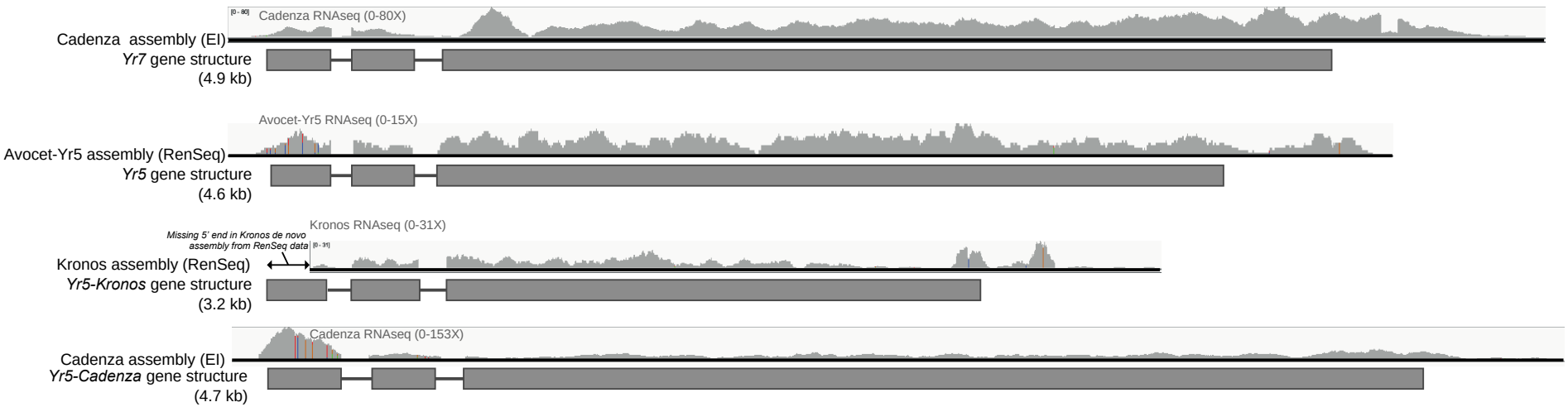


Figure 3-12. Comparison and validation of expression and gene structure of *Yr5* Kronos and *Yr5* Cadenza.
Black lines represent part of the scaffold (Cadenza assembly EI, Appendix 8-6) or RenSeq assembly contig (Avocet-Yr5, Kronos, Appendix 8-5). Grey rectangles show the exons and grey lines the introns in accordance to the mapped RNA-seq reads. Jianping Zhang confirmed the start ATG and stop codon with 5' and 3' Rapid Amplification of cDNA Ends (RACE) PCR. There was a missing part of the 5' end in Kronos RenSeq assembly as compared to the EI assembly and that is similar to what we observed for Cadenza in Figure 3-2. The coverage range is showed above the graph representing the reads (e.g Cadenza RNAseq 0-80x meant coverage ranges from 0 to 80x in *Yr7*).
RNA-Seq data for Cadenza were published as part of this study and were collected at flag leaf stage without treatment, the data corresponding to Avocet-Yr5 were obtained from Dobron et al., 2016 are derived from leaves infected with a *Pst* isolate avirulent to *Yr5*. Finally, RNA-seq data corresponding to Kronos were retrieved from Pierce et al., 2014 and correspond to flag leaf stage without treatment. Coloured vertical lines on the read graph correspond to SNPs between the assembly and the mapped reads. Reads mapping more than one location with the same score were not filtered out, thus very similar regions will appear with SNPs.

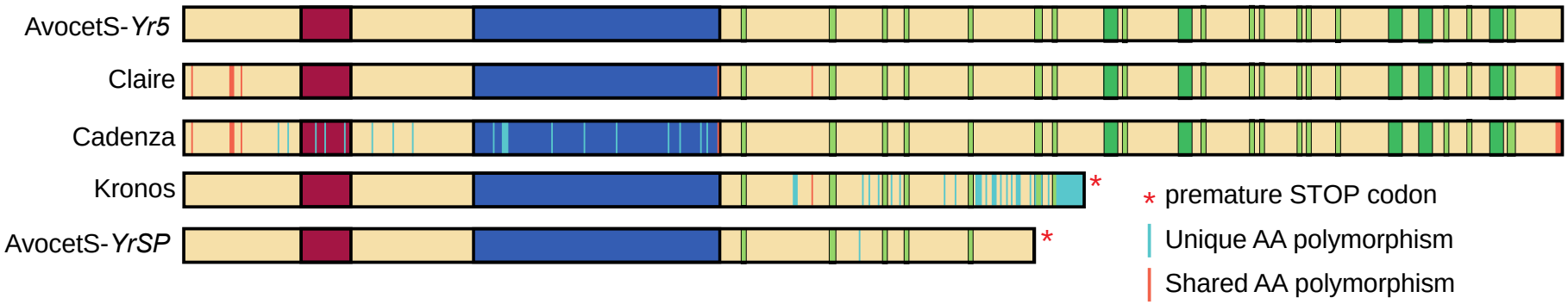


Figure 3-13. Five Yr5/YrSP haplotypes were identified in this study.
Polymorphisms are highlighted across the protein sequence with orange vertical bars for polymorphisms shared by at least two haplotypes and blue vertical bars for polymorphisms that are unique to the corresponding haplotype. Matching colours across protein structures illustrate 100% sequence conservation.

3.3.8. Developing diagnostic markers for *Yr7*, *Yr5* and *YrSP*

3.3.8.1. Designing diagnostic primers for Yr7

Based on an alignment of best blast hits found for *Yr7* in the genomes (described in Appendix 8-6 and above), we designed allele-specific markers for *Yr7*, *Yr5* and *YrSP* (Methods section 3.2.8). For *Yr7* we targeted polymorphisms located in exon 3, which is the most variable region, and designed 54 KASP primer assays. We tested all sets on a subset of the WAGTAIL panel (see section 3.2.8 for more details) and Figure 3-14 shows a representative subset of the results with positive and negative results. We had two different types of results overall (summarised in Table 3-10):

- Positive: the water samples did not amplify, there was a clear distinction between the two targeted alleles and all *Yr7* accessions were amplified for the *Yr7* allele (Yr7-A, Yr7-B and Yr7-D, 3/54 tested markers)

- Negative: water samples were amplified (Yr7-C and Yr7-F) or there was no clear distinction between the two targeted alleles (Yr7-F) or not all of the *Yr7* varieties were amplified for the *Yr7* allele.

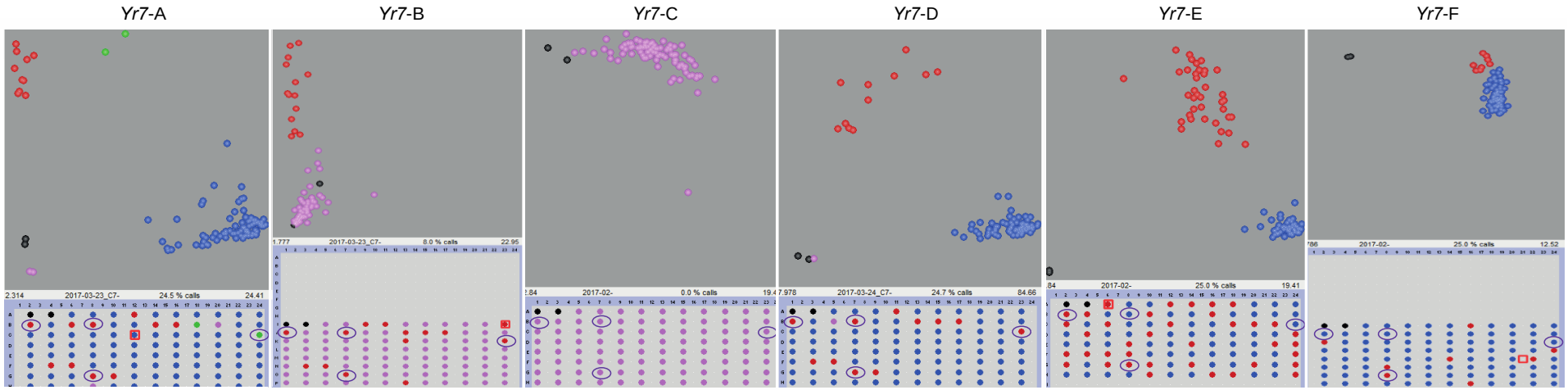
We could not differentiate between the water controls and the samples for the Yr7-C marker, as they both amplified in a similar manner (Figure 3-14). Thus, this marker was not suitable. There were a lot of samples amplifying for the *Yr7* allele when we used the Yr7-E marker, although only two out the four *Yr7* carriers were positive for the *Yr7* allele (red in Figure 3-14). Yr7-E was thus not specific to the targeted *Yr7* allele. For Yr7-F, it was very difficult to tell apart the signal from the VIC tail from the FAM tail. Moreover, the water controls were also amplified and not all the four *Yr7* carriers were detected

with the *Yr7*-specific marker (red). These three markers are thus not usable as diagnostic markers for *Yr7*.

All four *Yr7* carriers amplified for the *Yr7*-specific allele in Yr7-A and D assays (Figure 3-14). Additional samples amplified in a similar manner, although we cannot tell whether they are false positive or additional *Yr7* carriers we were unaware of. Yr7-B behaved like a dominant marker. Indeed, only samples that were positive for the *Yr7*-specific allele did amplify in this assay, including the four known *Yr7* carriers (Figure 3-14). Given that Yr7-A, B and D all detected the known *Yr7* carrier with a few additional samples, we only retained these three markers for the following test. Yr7-A and Yr7-D target specific SNPs between *Yr7* and its closest homologs in sequenced wheat genomes (Appendix 8-6) and Yr7-B targets an *Yr7*-specific insertion.

Table 3-10. Summary of the KASP assays carried out for *Yr7*

Tested marker	Results	Conclusion
Yr7-A	Clear separation between the two alleles Amplified all *Yr7* varieties + eight additional ones Water samples did not amplify	Suitable marker
Yr7-B	Only one allele amplified (dominant marker) Amplified all *Yr7* varieties + ten additional ones Water samples did not amplify	Suitable marker
Yr7-C	No separation between the two alleles Amplified water samples	Marker not suitable
Yr7-D	Clear separation between the two alleles Amplified all *Yr7* varieties + seven additional ones Water samples did not amplify	Suitable marker
Yr7-E	Clear separation between the two alleles Amplified all *Yr7* varieties + a lot of additional ones (> 20) Water samples did not amplify	Marker not suitable – could generate false positives
Yr7-F	No very clear separation between the two alleles Amplifies water samples	Marker not suitable



DNA plate layout

	1	2	3	4	5	6	7	8	9	10	11	12
A	Empty	Empty	AWARD	CONTENDER	AARDVARK	VPM1	WARRIOR**	ANGLO	HORNET	BILBO	DURIN	SABRE*
B	GRAFTON	HAMMER	AMBROSIA	TONIC	WALPOLE	DICKSON	VIRTUOSE	KWS_STERLING	SPARK	VICTO	NAPIER	ARGENT
C	RAGLAN	CADOGAN	SHANGO	DISPONENT	CARIBO	BOSTON	ORESTIS	DIABLO	BUSTER	GOLDLACE	GLADIATOR*	TOMMY
D	CROFTER	RIBAND	SANCERRE	BELUGA	ARISTOCRAT	GALAHAD	COCKPIT	ABELE	BRIMSTONE	MARIS_HUNT	SPRY	IMPALA
E	MANDATE	ELYSEE	SCOUT	MARIS_RANGER	TIMARU	CHIANTI	ATOU	ALCHEMIST	MAGELLAN	KINSMAN	KWS_PODIUM	PR21R60
F	FENDER	SOLEIL	RAFFLES	VUKA	SW_TATAROS	TIPSTAFF	HOLDFAST	POROS	MARIS_DOVE	ANDURIL	COURTOT	NEWMARKET
G	CELEBRATION	NSL_92-5719	ARLINGTON	BROCK	CYBER	DICKINS	VICTO	HEDGEHOG	HARROW	MARIS_ENVO	COMET	HUSSAR
H	KOGA_1	HURLEY	OBERON	OSTARA	ARK	ORTON	JACADI	REYDON	STEADFAST	SARSEN	PROFESSEUR_MARCHAL	OPTION

Figure 3-14. Illustration of *Yr7* KASP primer sets testing.
Graphical output from KlusterCaller from the *Yr7* KASP assays. Each circle represents a sample listed in the corresponding table below the graphs, which represent the DNA plate layout of the subset of the WAGTAIL panel we used for the test. The layout beneath each graph corresponds to the DNA plate layout. Red and blue colours show the signal for the VIC and FAM tails, respectively, with VIC being associated to the *Yr7* allele and FAM to the alternate allele. Pink shows DNA could not be assigned to one or the other of the alleles given that it amplified the same way as the water controls (black dots). Known *Yr7* cultivars are shown in purple on the table and with purple circles on the graphs. Red squares only represent the position of the cursor while displaying the graphs on the monitor and are thus not relevant for the analysis.

3.3.8.2. Testing the Yr7 markers on a set of Cadenza-derived varieties

We observed on Figure 3-14 that additional samples amplified for the *Yr7*-specific allele when tested with Yr7-A, B and D markers. To determine whether these could be false positive or actual *Yr7* carriers, we assembled a panel of Cadenza-derivatives that we tested against two *Pst* isolates that are avirulent to *Yr7* and one that is virulent to *Yr7* (Table 3-11). The rationale is that in the presence of *Yr7*, the variety's infection type should be 1 (resistant) for both *Yr7*-avirulent Pst isolates and 2 (susceptible) for the *Yr7*-virulent *Pst* isolate.

We used Vuka as a positive control for inoculation and absence of *Yr7*. The typical response of a *Yr7* carrier would thus be 1 – 1 – 2 in Table 3-11. However, we observed a 1 – 1 – 1 profile in Cadenza and this indicates that Cadenza carries resistance genes that are effective against the *Yr7*-virulent isolate. Thus, both 1 – 1 – 2 and 1 – 1 – 1 profiles could testify for the presence of *Yr7* in this study.

We can see in Table 3-11 that varieties that were positive for *Yr7* based on the Yr7-A, B and D markers had either one or the other profile. This suggests that none of these Yr7-positive lines based on the KASP assay was susceptible to a *Pst* isolate that is avirulent to *Yr7*. These results are thus consistent with our hypothesis that the Yr7-A, B and D markers are specific to *Yr7*. A few varieties (e.g Bennington, KWS-Kerrin, Brando) were susceptible to one of the two isolates avirulent to *Yr7* in addition to their susceptibility to the *Yr7*-virulent isolate. However, none of them carried the *Yr7* allele. There was a set of varieties displaying 1 – 1 – 1 and 1 – 1 – 2 profiles but were negative for the *Yr7* alleles. This set is assembled from Cadenza derivatives and Cadenza also displayed a 1 – 1 – 1 profile so it could be that other genes that were passed by Cadenza are resistant to all tested *Pst* isolates in a *Yr7*-independent manner.

Overall, these results are consistent with the hypothesis that Yr7-A, B and D markers are specific to the cloned *Yr7* allele (sequences available in (Appendix 8-8)). We thus concluded that Yr7-A, B and D were suitable for selecting for *Yr7* in wheat and used them to investigate the breeding history of *Yr7* in the UK (Figure 3-6) and its prevalence in four diversity panels (Appendix 8-9). It is nevertheless important to note that in the absence of allele sequence information, we could also select for other *Yr7* alleles that we do not know.

Table 3-11. Presence/absence of *Yr7* alleles in a selected panel of Cadenza-derivatives and associated responses to different *Pst* isolates (1: resistant, 2: susceptile).
Avirulent to *Yr7*: PST 15/151 and 08/21; virulent to *Yr7*: PST 14/106. Blue depicts the *Yr7* alleles and green the alternate alleles. Based on these results we added the classification of the variety (*Yr7*/non-Yr7) in the column on the right

				Yr7			
	Yr7 avirulent	*Yr7* avirulent	*Yr7* virulent	C	G	A	
Variety	WYR BLUE (15/151)	WYR NAVY (08/21)	WYR PURPLE (14/106	Yr7-A	Yr7-B	Yr7-D	
BROCK	1	1	2	C	G	A	Yr7
CADENZA	1	1	1	C	G	A	Yr7
CADENZA 2	1	1	1	C	G	A	Yr7
CAMP REMY	1	1	2	C	G	A	Yr7
CORDIALE	1	1	1	C	G	A	Yr7
CORDIALE 2	1	1	1	C	G	A	Yr7
CUBANITA	1	1	1	C	G	A	Yr7
GRAFTON	1	1	2	C	G	A	Yr7
GRAFTON 2	1	1	2	C	G	A	Yr7
KWS_STERLING 2	1	1	2	C	G	A	Yr7
KWS-CURLEW	1	1	1	C	G	A	Yr7
KWS-QUARTZ	1	1	2	C	G	A	Yr7
ORBIT	1	1	2	C	G	A	Yr7
PARAGON	NA	1	2	C	G	A	Yr7
RAFFLES	1	1	1	C	G	A	Yr7
SKYFALL	1	1	2	C	G	A	Yr7
SPARK	1	1	1	C	G	A	Yr7
TONIC	1	1	2	C	G	A	Yr7
TONIC 2	1	1	2	C	G	A	Yr7
AARDVARK 2	1	1	1	G	N-A	G	non-Yr7
ACROBAT	1	1	2	HET	N-A	G	non-Yr7
ARRIVA	1	1	1	G	N-A	G	non-Yr7
AXONA 2	1	1	2	G	N-A	N-A	non-Yr7
BANTAM	1	1	1	G	N-A	G	non-Yr7
BATTALION	1	1	1	HET	N-A	G	non-Yr7
BENNINGTON	2	1	2	G	N-A	G	non-Yr7
BOWINDO	1	1	1	G	N-A	G	non-Yr7
BRANDO	1	2	2	G	N-A	G	non-Yr7
CHABLIS	2	2	2	G	N-A	G	non-Yr7
CHOICE	1	1	1	G	N-A	G	non-Yr7
CODOGAN	1	2	2	G	N-A	G	non-Yr7
CONVOY	1	2	2	G	N-A	G	non-Yr7
COSTELLO	1	1	1	G	N-A	G	non-Yr7
CRUSOE	1	1	1	G	N-A	G	non-Yr7
DOVER	1	1	1	G	N-A	G	non-Yr7
DUNSTON	1	2	2	G	N-A	G	non-Yr7
DUXFORD	1	1	2	G	N-A	G	non-Yr7
EMERALD	1	1	2	G	N-A	G	non-Yr7
ENERGISE	2	2	2	G	N-A	G	non-Yr7
FREISTON	2	1	2	G	N-A	G	non-Yr7
GALLANT	1	1	2	G	N-A	G	non-Yr7
GALTIC	1	1	1	G	N-A	G	non-Yr7
GULLIVER	1	1	2	G	N-A	G	non-Yr7
HORATION	2	1	2	G	N-A	G	non-Yr7
HYPERION	1	2	2	G	N-A	G	non-Yr7
JORVIK	2	1	2	G	N-A	G	non-Yr7
KETCHUM	1	1	1	G	N-A	G	non-Yr7
KWS_SISKIN	1	1	1	G	N-A	G	non-Yr7
KWS_TRINITY	1	1	1	G	N-A	G	non-Yr7
KWS-BOHINEN	1	1	2	G	N-A	G	non-Yr7
KWS-HORIZON	1	1	1	G	N-A	G	non-Yr7
KWS-KERRIN	2	1	2	G	N-A	G	non-Yr7
KWS-KIELDER	2	2	2	G	N-A	G	non-Yr7
KWS-SANTIAGO	2	2	2	G	N-A	G	non-Yr7
KWS-SILVERSTONE	1	1	1	G	N-A	G	non-Yr7
LIMERICK	1	1	1	G	N-A	G	non-Yr7
MARIS DOVE	1	2	2	G	N-A	G	non-Yr7
MARKSMAN	1	1	1	G	N-A	G	non-Yr7
MOULTON	1	1	1	G	N-A	G	non-Yr7
PANORAMA	1	1	1	G	N-A	G	non-Yr7
REFLECTION	2	1	2	G	N-A	G	non-Yr7
REVELATION	1	1	2	G	N-A	G	non-Yr7
RGT_CONVERSION	2	1	2	G	N-A	G	non-Yr7
RGT_ILLUSTRIOUS	1	1	1	G	N-A	G	non-Yr7
RGT_SCRUMMAGE	2	1	2	G	N-A	G	non-Yr7
ROCKY	1	1	1	G	N-A	G	non-Yr7
SCANDIA	1	1	2	G	N-A	G	non-Yr7
SCORPION25	1	1	1	G	N-A	G	non-Yr7
SHIRAZ	1	1	2	G	N-A	N-A	non-Yr7
VELOCITY	1	1	1	G	N-A	G	non-Yr7
VUKA	2	2	2	G	N-A	G	non-Yr7
WARLOCK24	1	1	1	G	N-A	G	non-Yr7

3.3.8.3. Breeding history of Yr7 and prevalence in wheat diversity panels

Yr7 has been widely deployed in Europe and Australia in the 1970s[157]. It has been introgressed from durum wheat cultivar Iumillo (tetraploid) into Thatcher (hexaploid), which is a donor present in several pedigrees of modern elite varieties. Cadenza, a UK variety that is an important recurrent parent, carries *Yr7*. Our hypothesis is that we could retrace *Yr7* breeding history in the UK via investigating selected Thatcher descendants, including more modern varieties. Additionally, we could also determine its prevalence in older materials, including landraces, to determine whether other sources than Thatcher could also have been *Yr7* donors.

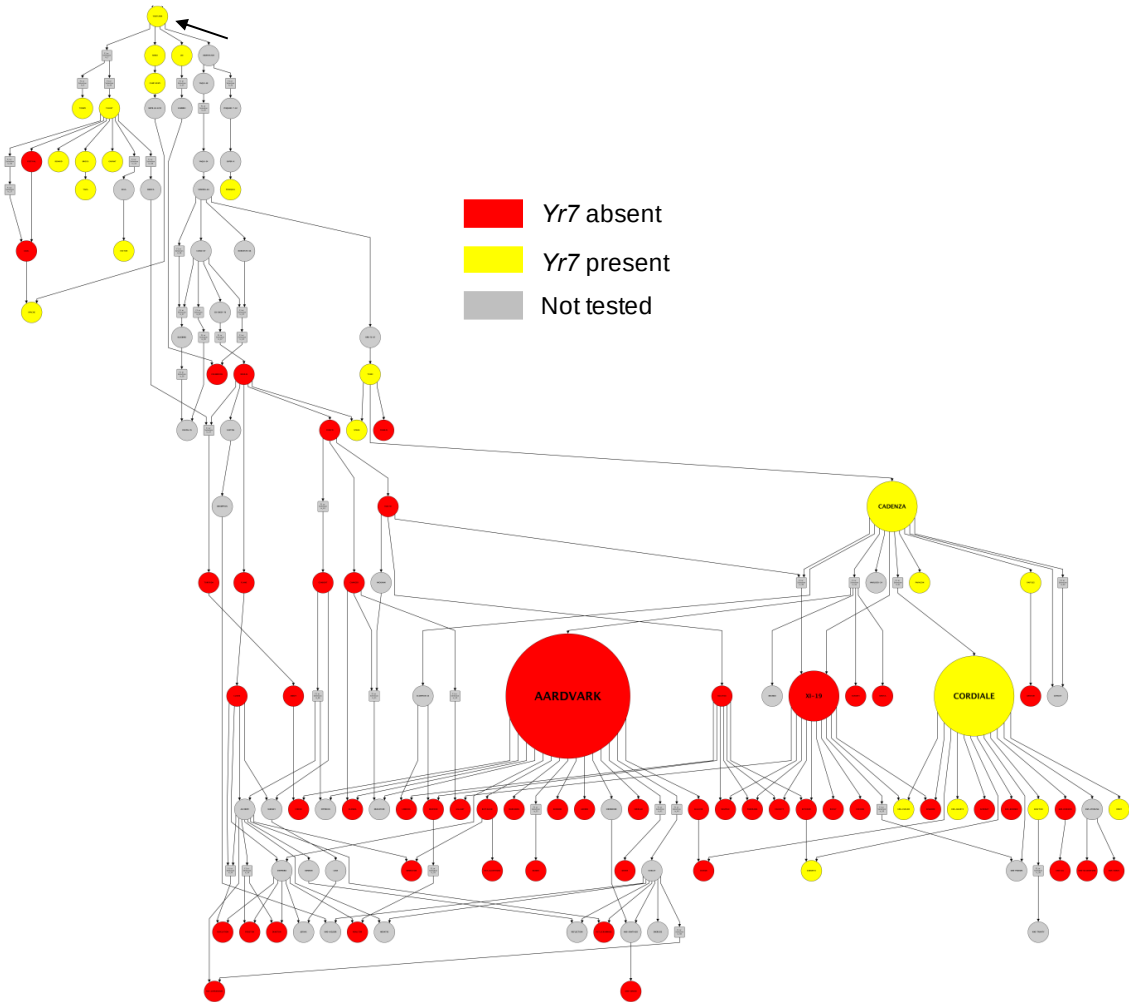


Figure 3-15. Pedigrees of selected Thatcher-derived cultivars and their *Yr7* status. Pedigree tree of Thatcher-derived cultivars where each circle represents a cultivar and the size of the circle is proportional to its prevalence in the tree. Colours illustrate the genotype with red showing the absence of *Yr7* and yellow its presence. Cultivars in grey were not tested or are intermediate crosses. *Yr7* originated from Triticum durum cv. Iumillo and was introgressed into hexaploid wheat through Thatcher (indicated by black arrow). Each *Yr7* positive cultivar is related to a parent that was also positive for *Yr7*. The figure was generated using the Helium software[208] (v1.17)

Breeding history of *Yr7*

We retrieved the pedigree data of the UK Cadenza-derivatives tested in Table 3-11 from the GRIS database (http://wheatpedigree.net/) and added Thatcher to determine whether *Yr7* was prevalent in UK breeding programs.

Yr7 is present both in historical and current UK varieties (Figure 3-15). *Yr7* has been widely deployed in Europe, Australia and New Zealand in the 1970s and the first virulence in Australia was recorded in 1986[157]. However, it is unlikely that it is currently actively selected for in the UK given that the resistance broke down approximately in 2010 (Chapter 2). One possible reason why *Yr7* is still present in more recent varieties is that other resistance genes and loci that are being actively selected for are linked to *Yr7*. For example, it has been shown that *Yr7* was linked with stem rust resistance locus *Sr9* (Iumillo carries *Sr9g*)[220], from which specific alleles confer resistance against the devastating *Pgt* race Ug99. Therefore, selecting for *Sr9g* might induce *Yr7* to be selected as well. However, in the case of the UK, there is no active selection for *Sr* genes yet. We thus suspect that *Yr7* is linked to another locus that is actively selected and this would be the reason why it is still present in more modern varieties.

Prevalence of *Yr7* in three characterised diversity panels:

We used the three *Yr7* KASP markers to genotype (i) cultivars from the AHDB Wheat Recommended List from 2005-2018 (https://cereals.ahdb.org.uk/varieties/ahdb-recommendedlists.aspx), representing recent elite varieties; (ii) the Gediflux collection of European bread wheat cultivars released between 1920 and 2010[5]; and (iii) the core Watkins collection, which includes older landraces from the 1900s (3.2.1).

(i) Results from the 2005-2018-UK_RL were consistent across already tested varieties: Cadenza, Cordiale, Cubanita, Grafton, and Skyfall were already positive. Energise, Freiston, Gallant, Oakley, and Revelation were negative on both panels as well. Results were thus consistent across different sources of DNA. *Yr7*-containing varieties are not prevalent in the 2005-2018 Recommended List set. This gene is present in Skyfall, which represents 11 % of the total UK acreage (Appendix 8-1).

(ii) The frequency of *Yr7* was relatively low in the Gediflux panel (4 %). This is consistent with results in (Figure 2-1): *Yr7* deployment started in the UK in 1992 with Cadenza and it was rarely used prior to that date.

(iii) We observed that 10 % of the core-Watkins collection was positive for the *Yr7*-specific allele (Appendix 8-9). All positive varieties originated from India and the Mediterranean basin. We know *Yr7* was introgressed into Thatcher (released in 1936) from tetraploid wheat Iumillo, which originated from Spain and North-Africa (Genetic Resources Information System for Wheat and Triticale - http://www.wheatpedigree.net/). Iumillo is likely to be pre-1920s. However, these landraces are all hexaploid wheat so they might have inherited *Yr7* from another source, although there is no evidence for *Yr7* coming from another source than Iumillo in the modern bread wheat varieties.

3.3.8.4. Developing gene-specific markers for Yr5 and YrSP

To our knowledge, *Yr5* has not been widely deployed in any region of the world. Currently, the University of California Davis (UC Davis) breeding programme is deploying *Yr5* in combination with *Yr15* as introgressions into modern elite cultivars. Germplasm derived from UC Davis is now also being used elsewhere (PAU, India).

Thus, we cannot carry out the same analysis as for *Yr7* and retrace its breeding history given the limited independent deployment of *Yr5*. The same is observed for *YrSP*. However, we know that *Yr5* comes from hexaploid spelt wheat and we can thus use this donor along with the introgression lines from UC Davis to test our *Yr5* marker.

We exploited the insertion located at the C terminus of *Yr5* when compared to its closest alleles in sequenced wheat genomes (Figure 3-13, Figure 3-16) to design a KASP marker that would select for the functional Yr5 protein we have evidence for. Additionally, we identified one SNP specific to *YrSP* in the allelic series and targeted it for *YrSP*-specific markers (Figure 3-17). We tested our *Yr5* marker on *Yr5* and *YrSP* donors spelt cultivar Album and wheat variety Spaldings Prolific, respectively, and on *Yr5+Yr15* (labelled 515) introgressed breeding material from UC Davis (Yecora Rojo 515, Redwin 515, UC 1745 515 and Summit 515) (Figure 3-16, Table 3-12). Only these introgressed lines and our *Yr5* positive controls amplified the signal corresponding to the *Yr5* allele (red on Figure 3-16). *YrSP* and other *Yr5* alternate alleles were amplified in the corresponding varieties AvocetS-YrSP, Spaldings Prolific, Cadenza, Claire and Paragon. No amplification was observed for our negative controls, showing that no *Yr5* allele was present in our tested *Yr7* varieties or in Lemhi. This confirmed that our marker could discriminate between *Yr5* known alternate alleles, *YrSP*, and the known functional *Yr5* allele derived from spelt wheat (Spelt-Yr5). We thus investigated whether the known functional *Yr5* was found in the Watkins landrace panel. However, no amplification was observed for Spelt-Yr5 and only *Yr5* alternate alleles were identified in 35.4 % of the Watkins panel (1,069 varieties) (data not shown).

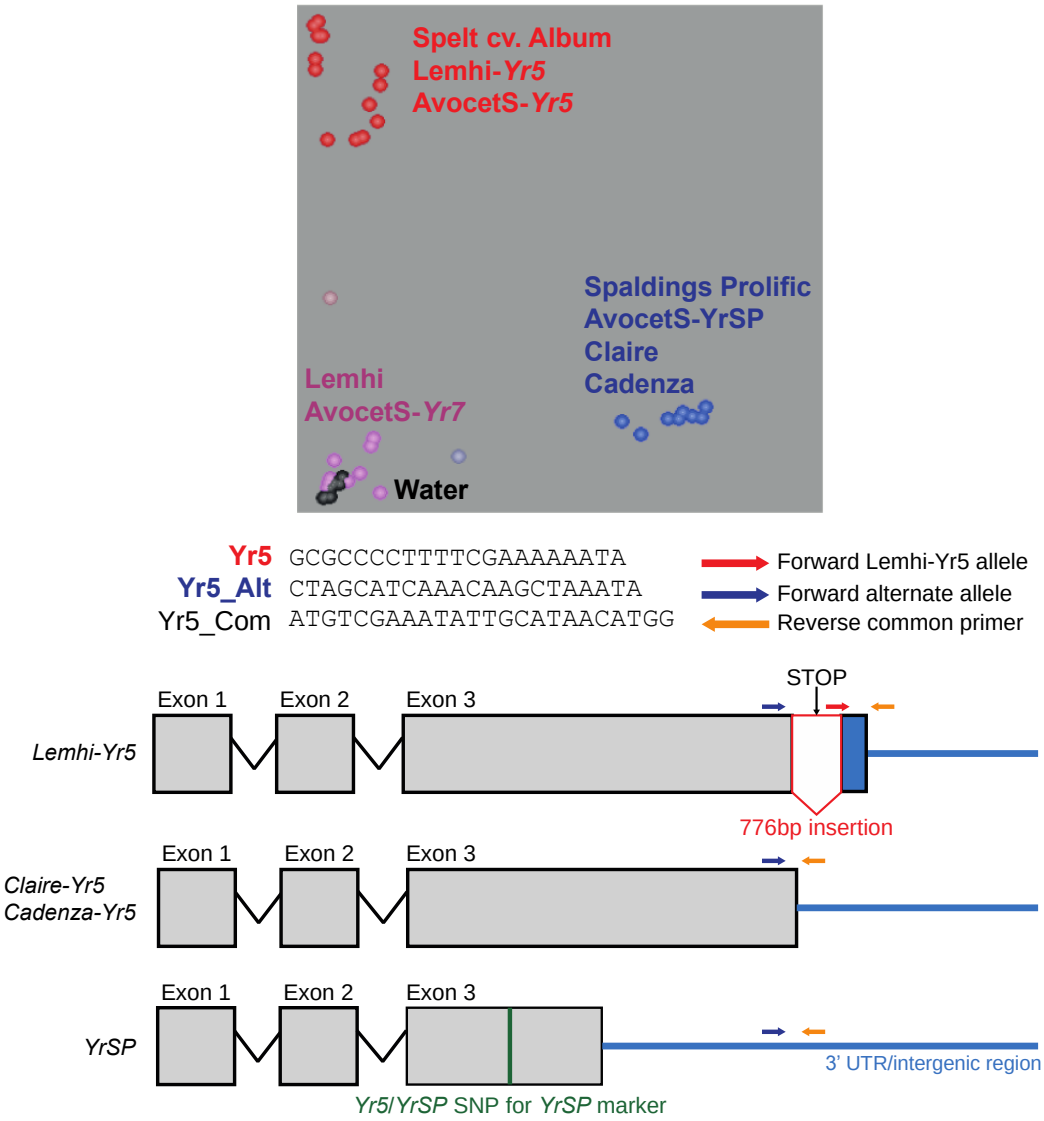


Figure 3-16. Illustration of *Yr5* KASP assay and schematics showing how we designed it.

Top: Graphical output from KlusterCaller from the *Yr5* KASP assays. Each circle represents a sample listed in Table 3-12. Red and blue colours show the signal for the VIC and FAM tails, respectively, with the corresponding primer sequence (without the tail) below. Pink shows DNA that did not amplify for the *Yr5* marker and black shows water control. Controls cultivars are shown in the matching colour with the amplified signal.

Bottom: Illustration of the primer positions on the *Yr5* allelic series. Blue/Orange primer pair only amplifies in *Yr5* alternate alleles and *YrSP* because of the 776bp insertion specific to the *Yr5* allele. Indeed, there is no extension step in the PCR program for KASP assay, thus this long fragment cannot be amplified, leading to the amplification of the Red/Orange fragment in *Yr5* only.

Table 3-12. Presence/absence of *Yr5* alleles in selected varieties.
We tested the KASP marker on the *Yr5* and *YrSP* donors spelt cultivar Album and Spaldings Prolific, respectively. We further tested the marker on Yr5-introgressed lines in AvocetS and Lemhi backgrounds and breeding material from the University of California, Davis breeding program (Yecora Rojo 515, Redwin 515, UC 1745 515 and Summit 515). We included bread wheat cultivars Claire, Cadenza and Paragon in which we identified alternate alleles for Yr5 (Figure 3-13). We used Iumillo (*Yr7* donor), Marquillo (Marquis x Iumillo), Lemhi and AvocetS-Yr7 as negative controls.

Variety	*Yr5*	*Yr5* alternative alleles	No amplification
Spelt cv. Album	Yes	-	-
AvocetS-Yr5	Yes	-	-
Lemhi-Yr5	Yes	-	-
Spaldings Prolifc	-	Yes	-
AvocetS-YrSP	-	Yes	-
Claire	-	Yes	-
Cadenza	-	Yes	-
Paragon	-	Yes	-
Yecora Rojo 515	Yes	-	-
Redwin 515	Yes	-	-
UC 1745 515	Yes	-	-
Summit 515	Yes	-	-
Iumillo	-	-	Yes
Marquillo	-	-	Yes
Lemhi	-	-	Yes
AvocetS-Yr7	-	-	Yes

For *YrSP*, we designed a primer set for KASP assays targeting the unique SNP between *YrSP* and the other *Yr5* alleles (Figure 3-17). We first tested the marker on *YrSP* donor Spaldings Prolific and AvocetS-YrSP, that are the only *YrSP* carriers we know of. We used varieties carrying the different *Yr5* haplotypes as negative controls (spelt 'Album', AvocetS-Yr5, Claire, Cadenza). The *YrSP*-specific allele was only amplified in Spaldings Prolific and not in the other varieties, confirming the specificity of our marker according to the tested material (Figure 3-17). Additionally, we tested *YrSP* marker on

the 2005-2018 UK Recommended List panel and *YrSP* was only present in the control AvocetS-YrSP (Appendix 8-9).

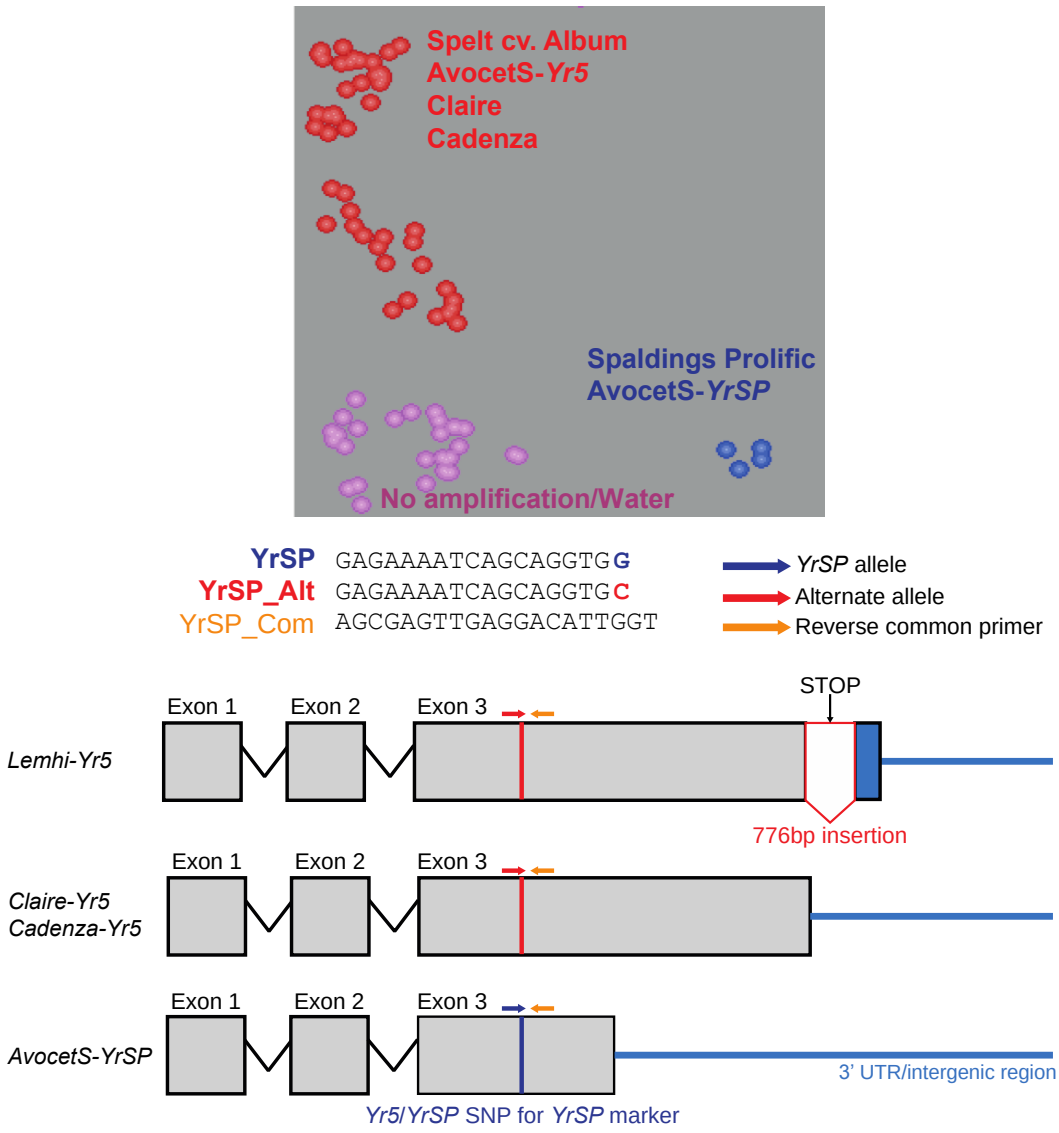


Figure 3-17. Illustration of *YrSP* KASP assay and schematics showing how we designed it.
Top: Graphical output from KlusterCaller from the *YrSP* KASP assays. Each circle represents a sample listed in Appendix 8-9. Red and blue colours show the signal for the VIC and FAM tails, respectively, with the corresponding primer sequence (without the tail) below. Pink shows DNA that did not amplify for the *YrSP* marker and black shows water control. Controls cultivars are shown in the matching colour with the amplified signal.
Bottom: schematics illustrating the position of the *YrSP*-specific SNP we identified between *YrSP* and the other *Yr5* alleles. The Blue/Orange pair will thus only amplify in presence of *YrSP* and the Red/Orange pair in the presence of the other alleles.

3.4. Discussion

We demonstrated in this Chapter that we successfully cloned *Yr7*, *Yr5* and *YrSP* using MutRenSeq. We confirmed the genetic position of the genes in F_2 populations and validated the presence of the cloned genes in the original donors durum wheat Iumillo, spelt wheat Album and bread wheat Spaldings Prolific, respectively. We generated 'perfect' markers to assist marker-assisted selection in breeding programs and tested them in two characterised diversity panel as well as a subset of the 2005-2018 UK Recommended List varieties.

3.4.1. MutRenSeq is a suitable approach to clone NLR resistance genes

Previous work used genetic mapping approaches to narrow down the *Yr7*, *Yr5* and *YrSP* genetic interval on chromosome 2B (Chapter 2). However even with the new chromosome-scale reference assembly available, the physical interval we defined based on all linked markers is still very wide (~ 150 Mb) and includes ~ 20 NLR loci in Chinese Spring, including some NLR clusters (Chapter 4). Thus, although genetic mapping has successfully been used in the past to clone resistance genes[100], it is often not sufficient on its own to provide a high enough resolution allowing the identification of the causal gene. Generating loss of function mutants and sequencing their NLR complement thus allowed us to uncover the genes carrying the causal mutations in all three cases.

MutRenSeq was successfully used before to clone *Sr22* and *Sr45*[137]. It is a powerful technique for wheat, where whole genome re-sequencing is still expensive if data from several lines and at a suitable coverage to call SNPs with confidence are required. However, it relies on a strong assumption regarding the nature of the targeted gene. Indeed, the target has to belong to the NLR family given that the capture array will only

hybridise NLR-related sequences. This is only possible because all NLRs share a very characteristic multi-domain structure that is specific to this protein class. In our case, *Yr7*, *Yr5*, *YrSP* and *Yr12* all display race-specific resistance, which is a hallmark of NLR-mediated resistance (Chapter 2). When no prior knowledge about the gene class is known, whole exome capture and sequencing of knock-out mutants could be a suitable alternative as here most genes, regardless of their type, are included in the capture platform. In our case, chromosome flow sorting and sequencing could also be used as we knew the chromosome location of the targeted genes[134].

MutRenSeq is a reference-free approach, which is very convenient when focusing on a gene family where presence/absence variation of alleles, and even whole loci, is high between varieties. Indeed, the reference Chinese Spring[112] does not carry any of our targeted genes so it would have been impossible to use it as a reference to map reads derived from our mutant lines. When MutRenSeq was developed, it was shown that RenSeq data were suitable to generate a draft *de novo* assembly of the NLR gene set in the captured sample and that this draft assembly was appropriate for read mapping and SNP calling[137]. However, we reported in Figure 3-2 that we were missing the 5' end of the *Yr7* candidate, which precluded the identification for one of the seven mutant lines. We thus used the available Cadenza sequence to correct the candidate contig, as the corresponding sequence was not captured. In the original MutRenSeq paper, a similar issue was flagged, but in this case the authors could retrieve the missing part in the assembly[137]. We identified the BED domain in this region in Figure 3-6. The Triticeae bait library does not include non-canonical NLR domains in its design so they are prone to be missed, especially when located at the extremities of an NLR. However, because *Yr5* and *YrSP* BED domains were successfully captured, we suspect that the library preparation must have accounted for this. Indeed, the average insert size was 374 bp in

the Cadenza wild-type reads, whereas it was 521 bp in Lemhi-Yr5 data, although read length was the same in both case (250 bp).

It is important to note that RNA-Seq data derived from varieties carrying *Yr7*, *Yr5* and *YrSP* were required to validate the gene structure of the candidates. Such data are not necessarily available on public repositories, especially if the genes are poorly studied. In our case, we could use previously published data from an *Pst* infection time course on AvocetS-Yr5[197] to validate *Yr5* and *YrSP* gene structure. We generated RNA-Seq data from Cadenza for *Yr7*. It is thus relevant to bear in mind that additional data might be needed to further study and validate the candidates identified with MutRenSeq.

The importance of confirming the gene structure of the candidate genes will depend on the objective. Do we need to know the structure of the causal gene to pursue functional characterisation? Do we need to know the structure to be able to design gene-specific markers for marker-assisted selection? In the first case, knowing the gene structure will be mandatory (discussed below), whereas it is optional in the second case. Indeed, one can design several markers targeting the candidate and test them in wider diversity panels, as we did for *Yr7*, *Yr5* and *YrSP*, without knowing the gene structure.

In some cases, the identified candidates could be homologous to already annotated NLRs and thus their structure can be predicted without additional RNA-Seq data. Indeed, providing the similarity between the candidate and its homolog is high enough, one can use the gene structure of the characterised one and transfer it onto the candidate. However, RNA-Seq will ultimately provide the confirmation. We think it is advised to find out whether RNA-Seq data from a variety carrying the gene of interest are available,

for example on public databases such as www.wheat-expression.com[108,221], or generate this data for the variety used for MutRenSeq.

Confirming the gene structure is crucial for:

- Determining whether the mutations have an effect on the candidate predicted protein, thus adding more evidence for the candidate to be the causal gene or not
- Functional validation in a susceptible background via transformation. The gene structure will be required to ensure the complete coding region with introns is used to design the constructs.
- Pursuing any further functional characterisation on the target gene

3.4.2. Validation of *Yr7*, *Yr5* and *YrSP* candidates

We demonstrated in Chapter 2 that the likelihood for seven mutant lines to carry mutations in the same gene by chance was extremely low ($P < 6.7E-8$). Even in the case of *YrSP*, where only four mutants were identified, the candidate contig was the only one identified by the MutantHunter program when setting the parameters retaining candidate(s) only if each mutant line carries a mutation in the bespoke contig (Jianping Zhang, personal communication).

We found the *Yr5* candidate only in spelt cultivar Album, AvocetS-Yr5 and Lemhi-Yr5 (Appendix 8-9, Figure 3-16) and the *YrSP* candidate only in AvocetS-YrSP and Spaldings prolific (Appendix 8-9, Figure 3-17). Additionally, each candidate was only found in the corresponding AvocetS-Yr line and all were absent in AvocetS.

We established the genetic linkage between the *Yr7* physical locus defined on RefSeqv1.0 and the *Yr7* candidate with traditional genetic mapping and bulked segregant analysis followed by exome capture and sequencing in two distinct F_2 population derived

from two independent *Yr7* knock-out mutants (Figure 3-8, Table 3-8, Figure 3-7, respectively). However, when we investigated the segregation of the resistant and susceptible phenotypes in F_2 progenies, the results were different from the expected 3:1 (Table 3-2). We attributed this to the significant number of plants we could not screen partly because of failed inoculation and developmental issue with some individuals. We used F_2 families derived from a cross between the M_7 mutant lines and the wild-type parent for the bulk segregant analysis, thus the majority of background mutations were fixed and this could have deleterious effects on germination and development. Crossing selected homozygous M_2 lines for the causal mutation to the wild-type could have prevented this. Alternatively, *Pst* inoculation success strikingly varies depending on the inoculation method[222] so testing an alternative method could have also helped. We confirmed the genetic linkage between *Yr5* and *YrSP* and markers located in the physical locus on RefSeqv1.0 with genetic mapping in the corresponding bi-parental populations (Figure 3-8).

The most common way to validate resistance genes in plants is to transform a susceptible variety with the gene of interest to determine whether it can provide resistance. This was successfully done for *Yr36*[63] and *Yr15*[146], among other cloned rust resistance genes. In the case of *Yr7, Yr5* and *YrSP*, we validated the candidates with confirming the genetic linkage between the candidate and the *Yr* locus and using allelic variation to design gene-specific primers and test them in characterised diversity panels. Additionally, we discussed above that the probability of having a contig carrying a mutation in seven lines by chance was extremely low ($P < 6.7E-8$). Despite this strong evidence, we initiated the generation transgenic plants for *Yr7, Yr5* and *YrSP* and this will be discussed in Chapter 5.

3.4.3. Elucidating the relationship between *Yr7*, *Yr5* and *YrSP*

First record of work carried out on the genetic linkage between *Yr7* and *Yr5* goes back to the 1980s[159]. Later work concluded on their allelic relationship as no recombinant could be found between crosses of *Yr7* and *Yr5* varieties[55]. *YrSP* was also described as allelic or closely linked to *Yr7* and *Yr5*[180]. Knowing whether these genes are true alleles or three different loci is important for breeders to select the right gene/allele and for the possibility to generate combinations of them. However, it is important to note that if the genes are physically very close to each other, they could be linked in repulsion. This means that the two functional variants we cloned do no coexist on the same chromosome and that no crossing-over would occur between the two loci. Thus, crossing two varieties carrying the desired *Yr7* and *Yr5* variant would not result in a progeny carrying the two variants (Figure 3-18).

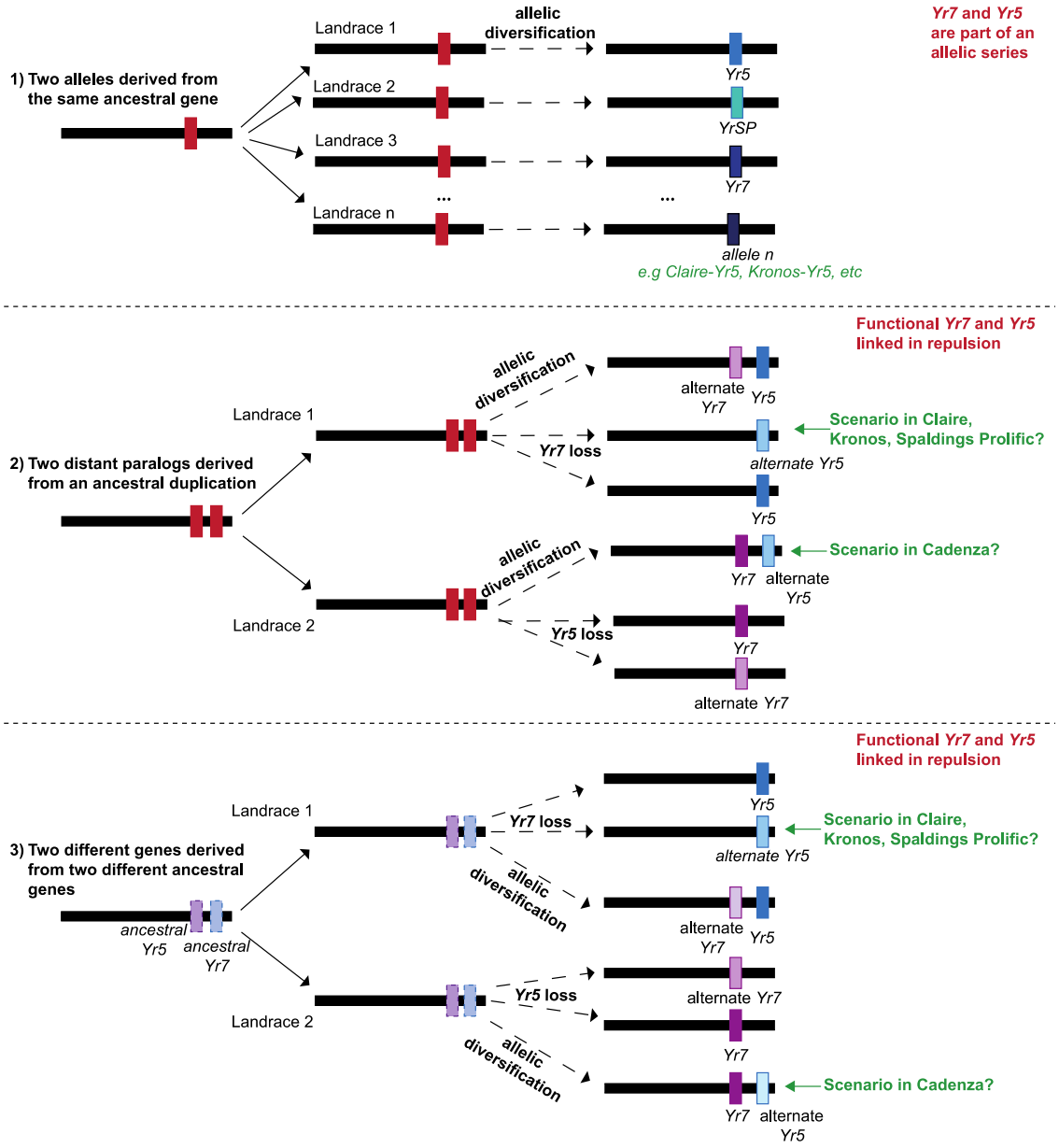


Figure 3-18. Schematics showing the different possible scenarios regarding *Yr5* and *Yr7* evolution and relationship.
1) *Yr7* and *Yr5* are two alleles of the same allelic series and are derived from the same ancestral gene.
2) *Yr7* and *Yr5* are derived from an ancestral duplication followed by allelic diversification and the two functional alleles we cloned are linked in repulsion.
3) *Yr7* and *Yr5* are derived from two different ancestral genes followed by allelic diversification and the two functional alleles are linked in repulsion. We included in green the possible scenarios for the *Yr5* alleles we identified in Figure 3-9

Our results show evidence that *Yr7* and *Yr5* are two different genes and that *Yr5* is allelic to *YrSP* (Figure 3-6 and Figure 3-18, scenario 3)). However, based on sequence only and allelism tests showing that no recombinant could be found between the three loci, we cannot discard the alternative explanations that *Yr5* and *Yr7* are closely linked paralogous genes that arose from a very recent duplication event (Figure 3-18, scenario 2) or that *Yr7* is an allele of *Yr5* that originated from a very diverse haplotype (Figure 3-18, scenario 1). The absence of recombination between the pairwise populations suggests that *Yr7*, *Yr5*, and *YrSP* are linked in repulsion, but we cannot discriminate between paralogous or allelic relationships. However, the high sequence identity alongside the genetic analyses support the hypothesis that *Yr5* and *YrSP* are derived from a common sequence and most likely constitute alleles, whereas *Yr7* is encoded by a closely related, yet distinct gene. Additionally, we identified a potential *Yr5* allele in the Cadenza genome. This allele is more divergent than the Claire allele (Figure 3-13, Table 3-9) so it is unclear whether it is a true allele of the same series. However, if that is the case, it would provide further evidence in favour of *Yr7* and *Yr5* being two different genes because Cadenza does carry *Yr7* (Figure 3-18, scenario 3).

3.4.4. Combining available wheat genome sequences enables designing gene-specific markers for *Yr7*, *Yr5* and *YrSP* and testing them in characterised wheat diversity panels

We developed gene-specific markers for *Yr7*, *Yr5* and *YrSP* for marker-assisted selection in breeding programs and tested them on different panels including (i) a set of potential *Yr7* carriers based on literature research (*Yr7* marker only), (ii) a set of varieties that belonged to the UK AHDB Recommended List (https://cereals.ahdb.org.uk/varieties/ahdb-recommended-lists.aspx) between 2005 and

2018 (labelled 2005-2018-UK_RL, *Yr7* and *YrSP* markers), (iii) the Gediflux collection that includes modern European bread wheat varieties (1920-2010)[191] (*Yr7* marker), (iv) a core set of the Watkins collection, which represent a set of global bread wheat landraces collected in the 1920-30s[190] (*Yr7* and *Yr5* markers)

Yr7, *Yr5* and *YrSP* positive alleles were identified in all their respective donors (Figure 3-14, Figure 3-16, Figure 3-17). We tested our *Yr5* marker on a set of Yr5+Yr15 introgressions lines from UC Davis and all were positive for the cloned *Yr5* allele, further confirming the specificity of the marker (Table 3-12, Figure 3-16). We carried out a similar analysis with our *YrSP* marker and showed that only Spaldings Prolific and AvocetS-*YrSP* were positive for the cloned allele (Figure 3-17). When tested on a set of varieties from the AHDB recommended list (2005-2018), none of the varieties were positive for YrSP, which is consistent with our knowledge that *YrSP* has never been deployed in the UK (Appendix 8-9).

When testing our *Yr7* marker on a panel of Cadenza-derivatives, there was a good correlation between an expected phenotype for *Yr7* and the presence of the *Yr7* allele (Table 3-11). No variety susceptible to all isolate was positive for the *Yr7* allele, which is consistent with our expected Infection Type for *Yr7*. However, a significant amount of varieties displaying a Cadenza-like response were negative for the *Yr7* allele. This can be due to the fact that the varieties carry additional resistance genes effective against the tested *Pst* isolate that we are unaware of. Indeed, we observed differences between the response to the two *Yr7*-avirulent isolates tested, showing that the cultivar background plays an important role.

We identified *Yr7* in a small proportion of the Gediflux and Watkins panels (4.3 and 9.9 %, respectively, Appendix 8-9). This is consistent with *Yr7* deployment history: *Yr7* was introgressed into the variety Thatcher (released in 1936) from tetraploid wheat Iumillo but was not widely deployed in the UK before 1992 when Cadenza was released and even at this date, Cadenza was not extensively cultivated (Chapter 2). Given that the Gediflux panel represents European varieties from 1920s – early 2000, it is thus expected that only a small proportion of varieties carries it. Based on the results from the Gediflux panel, *Yr7* was not only present in UK varieties but also several French and Dutch cultivars. All have Thatcher in their pedigree (http://wheatpedigree.net/), thus it is consistent with the above statement.

The Watkins panels is composed of a set of global bread wheat landraces collected in the 1920-30s so it predates *Yr7*'s introgression in Thatcher from tetraploid Iumillo and no landrace from England was positive for the *Yr7* allele. Only eleven landraces were positive and they were collected from the Mediterranean basin and India. However, all eleven are hexaploid wheat and, to our knowledge, Thatcher is the only *Yr7* source of all known *Yr7*-varieties. It is thus likely that these eleven landraces might have inherited *Yr7* from another source

Only five out of the set of 113 varieties that were on the AHDB recommended list between 2005 and 2018 carry *Yr7*: Skyfall, Cordiale, Cubanita, Ruskin and Grafton. Cubanita and Ruskin were never extensively cultivated between 1990 and 2016 (Appendix 8-1), Cordiale is cultivated since 2003 with a peak in 2009 (7.1 % of total harvested wheat) followed by a decreased since then and a similar pattern is observed for Grafton, which is cultivated since 2008 and peaked in 2012 (5.7 % of total harvested wheat). Both varieties thus started to decline when the prevalence of *Yr7* virulence in

tested *Pst* isolates increased in the field (reached 100% in 2012, Figure 2-1). Skyfall carries *Yr7* and was a widely grown variety in 2016 (11% of total harvested wheat) and despite its susceptibility to most of tested *Pst* isolates (https://ahdb.org.uk/ukcpvs), it performs well on the field overall and is still part of the 2019/2020 Recommended list (https://ahdb.org.uk/rl). With the *Yr*7 markers, we could thus make a direct link between the data we presented in Chapter 2 and the cloned *Yr7* allele. This provide valuable knowledge to monitor the efficiency of a gene in the field and identify establish a correlation between the gene and the prevalence of *Pst* isolate virulent to this gene at any given time.

Overall, we can be confident that our markers are specific to the cloned alleles of *Yr7*, *Yr5* and *YrSP*. However, it is important to bear in mind that we designed these markers based on available sequences at the date of the study. When more information becomes available we may identify additional alleles and this will allow us to improve the markers. Moreover, we already identified alternate alleles for *Yr5*. Although to our knowledge they do not provide the same resistance as Spelt-Yr5, it would be useful to determine whether they are functional.

Indeed, provided the alternative *Yr5* alleles are all functional against different *Pst* isolates, this would consist of a portfolio of alleles that could be used in breeding programs. It would be important to determine which polymorphism are actually mediating the resistance spectra, similarly to the work that has been done on the *Pik* alleles in rice (Josephine Maidment, John Innes Centre). Indeed, coupling this knowledge with the on-going development of gene editing techniques and their direct use in elite cultivars directly[223] would thus allow engineering potential new alleles and deploying them in the field in a shorter timeframe than traditional breeding.

Additionally, we will discuss further in Chapter 6 why designing and testing diagnostic markers should be part of cloning resistance genes in wheat. This would allow a direct application of the research in breeding programs.

3.4.5. *Yr7*, *Yr5* and *YrSP* encode BED-NLRs

The predicted protein sequences from *Yr7*, *Yr5* and *YrSP* coding sequences showed that they encode non-canonical NLR proteins. In addition to the NB-ARC domain and LRR regions at the C-terminus, they carry a BED domain at the N-terminus of their amino-acid sequences.

This is the first evidence showing that the BED-NLR structure is functional against pathogenic fungi. This complements previous work that showed that this particular type of protein was also effective in rice against the bacterium *Xanthomonas oryzae* pv. *oryzae*[224,225], the causal agent of bacterial blight. More recently, a candidate gene for *Xo1*, which provides resistance against both bacterial blight and bacterial leaf streak (latter caused by *Xanthomonas oryzae* pv. *oryzicola*), was also found to encode a BED-NLR immune receptor[226]. Thus, this particular domain architecture is effective against both fungal and bacterial pathogens. However, little is known about their mode of action in rice. *Xa1* has been shown to be effective against a specific class of effectors, the transcription activator-like effectors (TAL) effectors[227] but the mechanisms of recognition itself is unknown.

BED-NLRs were identified in several plant genomes[87,228,229] and hence labelled as Integrated Domain NLRs (ID-NLRs)[230]. This nomenclature refers to the Integrated decoy model where the non-canonical domain, or integrated domain, plays the role of effector

trap and mimics the original target of the effector to competitively promote binding of the effector to the immune receptor to trigger defense response[211]. Three examples of such immune receptors have been well studied: *RRS1* in Arabidopsis[45,231] and *RGA5*[232] and *Pik* in rice[233]. Interestingly, *RRS1* is able to recognise effectors from both bacterial and fungal pathogens. However, in all cases the protein responsible for the recognition of the effect acted in pair with another NLR protein and both partners are required for triggering defense responses. Moreover, it was shown that the Heavy-Metal-Associated (HMA) integrated domain was the most variable region of the protein between *Pikm* alleles[234]. This is not the case for Yr7, Yr5 and YrSP, as we showed that there was only one amino-acid change between Yr7 and Yr5 BED domain and that the BED domain is actually in the most conserved region between the two proteins (Figure 3-9). Thus, it is unlikely, in our opinion, that the BED domain solely drives *Yr7*, *Yr5* and *YrSP* specificity. Either it could function in a similar way to the integrated WRKY domain in the *Arabidopsis* RRS1-R immune receptor, which binds unrelated effectors and yet activates defense response through mechanisms involving the integrated domain with other regions of the protein[231]. Alternatively, BED domain function could be unrelated to pathogen recognition *per se* and is involved in defense signalling. We will further investigate these hypotheses in Chapters 4 and 5.

The BED domain itself was first defined in 2000 and was shown to bind DNA[235]. In plants, there was one particular BED domain-containing protein family showed to bind DNA, the *daysleeper* family[236]. The same work showed that knocking-out this particular gene had deleterious effect on the development of *Arabidopsis* plants. Little more is known about this gene family however. Because *sleeper* genes have similarities with hAT transposases, it has been hypothesized they arose from neo-functionalization of a domesticated hAT transposase[237]. Such domestication events were documented for a

wide variety of transposases but not BED-containing hAT transposases specifically[238]. Further analyses will be required to determine which mechanism drove BED domain integration to an NLR protein.

3.4.6. Summary

We cloned *Yr7*, *Yr5* and *YrSP* and developed and tested diagnostic markers to assist their selection in breeding programs. The BED-NLR architecture of these proteins led us to three main hypotheses regarding the function of the BED domain: (a) the BED domain is involved in direct or indirect effector recognition and acts in a similar manner as described for the integrated domain model, or (b) given that BED sits in one of the most conserved regions between *Yr7* and *Yr5* and is identical between *Yr5* and *YrSP*, it is involved in signalling to trigger defense response in the presence of the pathogen. We carried out both comparative genomics and molecular biology analysis to address these hypotheses in Chapter 4 and Chapter 5.

4. Analysis of the *Yr7*, *Yr5* and *YrSP* locus in wheat and related species and characterisation of BED-NLRs in plant genomes

4.1. Introduction

We demonstrated in Chapter 3 that *Yr7* and *Yr5* encode BED-NLRs and this led us to generate two alternative hypotheses regarding the BED domain's function in defense response:

(a) the BED domain is involved in effector recognition (directly or indirectly) and acts in a similar manner as described for the integrated domain model.

(b) the BED domain is involved in signalling to trigger defense response in the presence of the pathogen.

In this Chapter, we will focus on hypothesis (a) and use a combination of comparative genomics and neighbour-network analyses to determine whether *Yr7* and *Yr5* show features that would favour their involvement in direct effector recognition via their BED domain.

4.1.1. The integrated decoy model

The integrated decoy model was first described in 2014[44], where the authors used the latest findings regarding the mode of action of two NLR pairs from rice and Arabidopsis RGA4/RGA5[239] and RPS4/RRS1[240] to propose a model explaining their function. In each NLR pair, one gene contained an additional domain that is not usually found in NLRs

(RATX1, or HMA, for RGA5 and WRKY for RRS1), whereas the other gene of the NLR pair was canonical and required to trigger cell death. In both cases, the additional domain was shown to directly interact with the corresponding effector. The interaction between the effector and the additional domain, or 'integrated domain', would subsequently lead to the activation of the defense response and provide resistance against the pathogen. The authors thus proposed that the integrated domain serves as a 'decoy' for the effector, mimicking the original target of the effector in the plant[44].

Note that the terminology 'integrated decoy domain' has been discussed by Wu et al., 2015[241]. The authors proposed to refer to the integrated decoy domains as 'sensor domains' given that it is unknown whether the integrated domains are 'true' decoys in that they have lost the biochemical activity of the original effector target[81]. However, for clarity purpose, in this thesis we will refer to this model as the 'integrated decoy model' as it was first described as such.

We will describe below three well-characterised NLR pairs and what is known of their mode of action to date. This will allow us to identify shared characteristics among NLR pairs that we can further investigate using comparative genomics. Figure 4-1 recapitulates the structure of each NLR of the pair carrying the integrated domain.

4.1.1.1. *RGA4/RGA5*

RGA4 and *RGA5* encode two rice Coiled-Coil-NLRs (or CNLs) that provide resistance against *Magnaporthe oryzae*[242]. RGA4 is a canonical CNL, whereas RGA5 carries an additional domain at the C-terminus, downstream of the LRR repeats[232]. This domain showed similarities with a heavy metal associated (HMA) domain related to the cytoplasmic copper chaperone ATX1 from *Saccharomyces cerevisiae* and was thus

called RATX1 (also known as HMA)[232]. Both NLRs are in closed proximity and in head-to-head orientation in the rice genome. Cesari et al., 2013[232] showed that both *RGA4* and *RGA5* are required for resistance against *M. oryzae* isolates carrying the effectors AVR-*Pia* and AVR-*Pi-CO39* and that the RATX1 domain of RGA5 was able to bind both AVR-Pia and AVR1-CO39 effectors.

Further work demonstrated that overexpression of *RGA4* alone in *Nicotiana benthamiana* led to cell-death and that the co-expression of *RGA5* with *RGA4* prevented this auto-immune response to occur[239]. RGA4 and RGA5 are able to form homo- and hetero-complexes through their Coiled-Coil domains[239]. Additionally, the interaction between RGA5 and AVR-Pia leads to the de-repression of RGA4, allowing cell-death to be triggered. Overall this provided more knowledge regarding the mode of action of NLR pairs: in the absence of the pathogen, RGA5 represses RGA4, thus preventing cell-death to constitutively occur. In the presence of the pathogen, the interaction between AVR-Pia and HMA domain of RGA5 leads to the de-repression of RGA4 and cell-death is subsequently triggered.

4.1.1.2. <u>*Pik-1/Pik-2*</u>

Pik-1 and *Pik-2* encode CC-NLRs in rice and also confer resistance against *M. oryzae*[243]. Similarly to *RGA4* and *RGA5*, both loci are close to each other in the genome and in head-to-head orientation[243]. Pik-1 carries a non-canonical HMA domain between the CC domain and the NB-ARC domain. Interestingly, Pik-1-HMA and RGA5-HMA share a common ancestor[239]. However, given that their positions in the corresponding NLR are different it is likely that they arose from independent integration events (Figure 4-1)[239].

Both *Pik-1* and *Pik-2* are required for resistance[243]. It was shown that AVR-Pik was the effector that the Pik-1/Pik-2 pair recognised[244]. Two alleles of this NLR pair have been characterised: *Pikp-1*/*Pikp-2* and *Pikm-1*/*Pikm-2*. Pikp-2 and Pikm-2 are nearly identical, whereas Pikp-1 and Pikm-1 are more divergent and the variation is concentrated in the integrated HMA domain[234].

Maqbool et al., 2015[233] demonstrated that the Pikp-1/Pikp-2 NLR pair was able to recognise several AVR-Pik alleles with different affinities via direct interaction between the Pikp-1 HMA domain and the effector. The authors identified specific residues that were crucial for this interaction and the subsequent activation of defense response[233]. De la Concepcion et al., 2018[234] dissected the allelic specificities of Pikp and Pikm pairs for the different AVR-Pik alleles and resolved the corresponding structure of the binding interfaces. The authors showed that Pik-HMA interfaces were very plastic and supported the differential recognition of AVR-Pik variants and found evidence for a threshold of binding that was necessary for the activation of defenses upon recognition[234].

4.1.1.3. <u>*RPS4/RRS1*</u>

RPS4 and *RRS1* genes encode TIR-NB-LRR proteins (or TNLs) in Arabidopsis and are both required for race-specific resistance to bacteria and to fungi[245,246]. This TNL pair is also located in close proximity in the genome. Both RRS1 and RPS4 proteins interact in part via their TIR domains; this interaction is essential for defense activation[240]. *RRS1* alleles (*RRS1-R* and *RRS1-S*) carry a non-canonical WRKY domain at the C-terminus of the protein. The RRS1-R (resistant) WRKY interacts with AvrRPS4 (from *P. syringae*) Type-III effector, whereas this interaction does not occur with RRS1-S[231]. Furthermore, although PopP2 (Type-III effector from *Ralstonia solanacearum*) binds and acetylates residues in the WRKY domains in both RRS1-S and RRS1-R, this acetylation triggers

defense response in RRS1-R, but not in RRS1-S[45,231]. Additionally, RRS1-R is able to trigger defense response in the presence of *Colletotrichum higginsianum*, although the corresponding effector is unknown[231]. Given that the most obvious difference between RRS1-S and RRS1-R is the presence of an additional 90 residues at the C-terminus of RRS1-R, the authors concluded that this region was important for defense signalling in response to PopP2 and AvrRps4[45,231].

WRKY domains are able to bind DNA. The PopP2 effector is able to acetylate other WRKY-containing proteins and this acetylation prevents DNA binding in *N. benthamiana*[45]. Two of these WRKY-containing proteins were shown to promote PAMP-induced MAPK (mitogen-activated protein kinase) signalling in Arabidopsis[247]. The authors thus concluded that PopP2 likely facilitates bacterial infection via impairing WRKY-mediated activation of defense response.

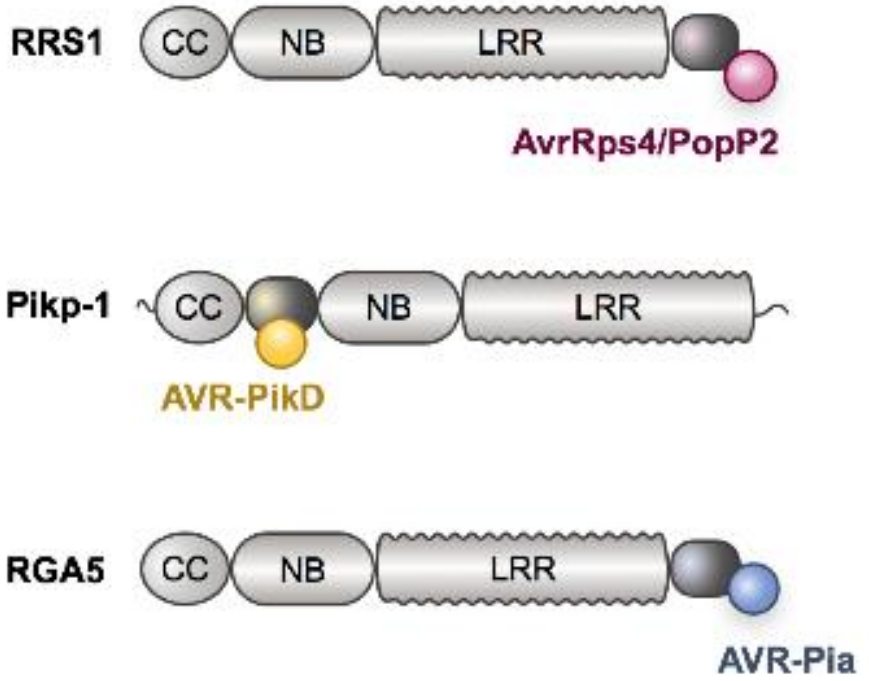


Figure 4-1. Schematics describing the current integrated decoy model in the three examples described above (figure adapted from Fujisaki et al., 2017[248]).
The NLR with the integrated domain (dark grey oval) directly interacts with the effector via the integrated domain and this direct recognition activates defense response.

We mentioned in Chapter 3 that BED domains are also able to bind DNA[249] and are present in the *daysleeper* transcription factor family in plants[236,237,250]. Knock-out mutants of *daysleeper* in Arabidopsis led to severe developmental effects[236]. Although it is not clear whether these genes are involved in disease resistance, it is tempting to speculate that these transcription factors could be the targets of *Pst* effectors that are recognised by BED-NLRs via a mechanism that is similar to *RRS1-R* recognition.

4.1.2. Whole genome studies to identify NLR-ID

Several studies showed that NLR carrying non-canonical domains, or integrated domains, are present in several plant genomes. Kroj et al., 2016[87] explored 31 proteomes from the Greenphyl database (https://www.greenphyl.org/cgi-bin/index.cgi) and found that among 2,699 canonical NLRs (NB-ARC and LRRs), 94 carried additional domains (3.5 % in 22/31 of the studied genomes). Additionally, Kinase, WRKY and BED domains were among the most frequent additional domains found in these NLR-IDs (11/94, 9/94 and 9/94, respectively[87]). Interestingly, BED domains were not found in TNLs. Given that the additional domains were located at different position within the NLR proteins, the authors concluded that multiple and independent integration events might have occurred during evolution of plant NLRs.

Bailey et al., 2018[88] carried out a similar study focused on grass genomes. Using phylogenetics, they determined that the majority of NLR-IDs were located in three separate clades in a phylogenetic tree containing all grass NLRs. Most NLR-IDs were found in one major clade (MIC1). Interestingly, most of the BED-NLRs grouped in their own clade (MIC3)[88]. A second clade (MIC2) grouped NLR carrying an integrated DDE domain (from the DDE endonuclease superfamily). These observations suggest that the backbone (here NB-ARC) of NLRs carrying BED domains are phylogenetically different

from NLR carrying most of the other additional domains. A similar trend was observed for NLR-DDE. The integration of the BED and DDE domain might thus be more ancestral than the domain integrations in the MIC1 clade, or alternatively the NLR backbone of MIC3 favours BED domain integration.

4.1.3. Beyond the integrated decoy model: indirect detection of pathogen effectors via integrated domains

The examples discussed above involve direct binding between the integrated domain and the effector. However, we hypothesized that BED-NLRs could also indirectly recognise effector. This hypothesis is derived from recent work on the *Pii-1* and *Pii-2* rice CNL pair that mediates *AVR-Pii M. oryzae* effector recognition[251]. In this pair, Pii-2 contains an integrated domain with a NOI motif in its C-terminal region. Fujisaki et al., 2017[248] showed that Pii-2 does not directly interact with AVR-Pii, but with OsExo70-F3, which is the target of AVR-Pii (Figure 4-2). The authors thus hypothesized that the Pii pair was guarding OsExo70-F3 and able to detect any alteration in this protein in response to its binding with AVR-Pii. This shows that integrated domains could also indirectly recognise effectors.

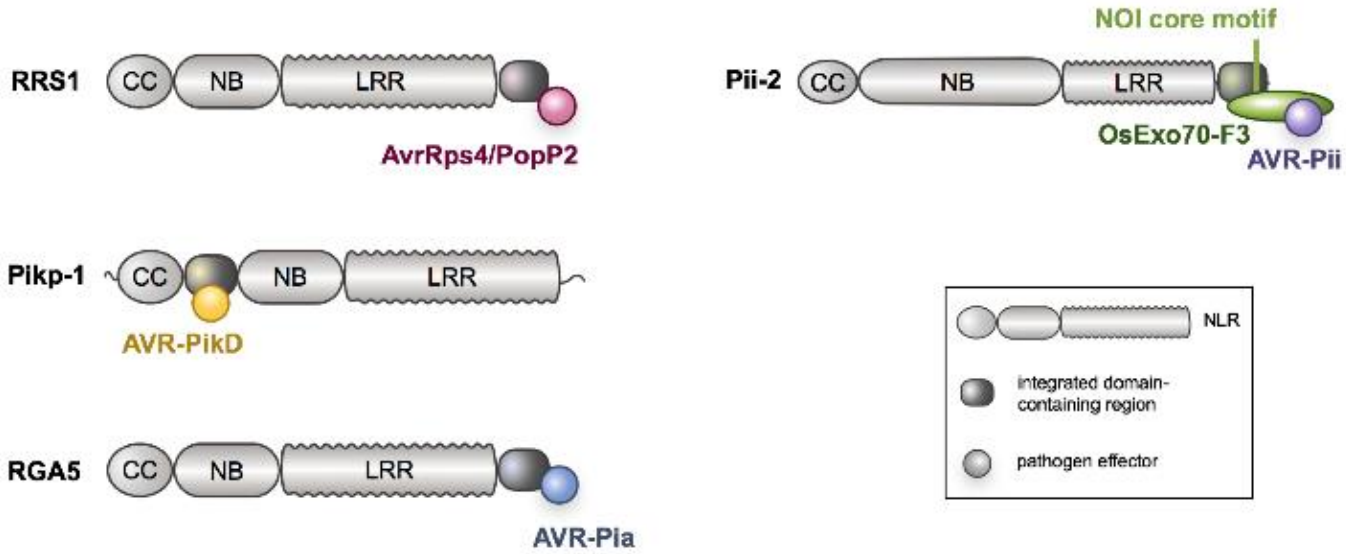


Figure 4-2. Schematics describing the current integrated decoy model in the three examples described above and including the recent evidence of indirect recognition via a NLR-ID (figure adapted from Fujisaki et al., 2017[248]).

Although numerous NLR-IDs have been identified in plant genomes, little is known about their mode of action, except from the very well characterised NLR pairs described above. The integrated decoy model proposes that the non-canonical domain is involved in effector recognition. In the discussed examples, the integrated domains were able to directly bind the corresponding effector. The PopP2 effector was also able to acetylate specific residues of the integrated WRKY in RRS1-R and RRS1-S, although this was not enough to trigger defense response in RRS1-S and suggested the involvement of other regions of RRS1-R in this process. AVR-Pii recognition by the Pii pair is dependent on the binding of OsExo70-F3 by Pii-2 once AVR-Pii interacted with OsExo70-F3. This shows that integrated domains are able to indirectly recognise effectors.

These two main hypotheses can be further explored using comparative genomics. From the different studies discussed above, we identified shared characteristics between functional NLR pair containing a partner with an integrated domain:

- The NLR pair is physically close and in head to head orientation.

- The integrated domain shares similarities with the same domain present in other proteins (e.g. PopP2 is able to bind and acetylate WRKY domains from both RRS1 and other specific WRKY containing proteins).

We did not find evidence of an NLR partner for *Yr7*, given that all identified loss of resistance mutants carried a mutation in *Yr7* gene. However, we cannot discard the hypothesis that a potential helper, that is functionally redundant in wheat, is required. Additionally, McGrann et al., 2014[174] originally found loss of resistance mutants in Lemhi-Yr5 mutant lines whose loss of resistance was complemented in the F_2 derived

from a cross to the susceptible variety Vuka. Vuka does not carry *Yr5* so the authors hypothesized that these mutants carried an independent mutation that was important for the *Yr5*-mediated resistance. Consequently, although we did not identify an NLR partner required in *Yr7* mediated resistance, exploring the *Yr* locus in a *Yr7* or *Yr5* carrier might uncover new features that we missed in the genomes we explored in Chapter 3.

Additionally, we hypothesized that investigating similarity between BED domains from BED-NLRs and BED domains from BED-containing proteins may provide leads regarding the initial effector target.

4.1.4. Summary

In this Chapter, we present how comparative genomics in the wheat pangenome allowed us to identify an additional *Yr7* allele and investigate the sequence and synteny conservation of the *Yr* locus in wheat and related grasses. Additionally, neighbour-network analyses on BED domains derived from BED-NLRs and other BED-containing proteins suggested that a BED domain from a specific BED-protein architecture was associated with BED-NLRs. Together these results enable us to clearly define hypotheses that will be tested in Chapter 5.

4.2. Material and Methods

4.2.1. *Yr7* and *Yr5* alleles identification in the wheat pangenome

We used the *Yr7* and *Yr5* sequences defined in Chapter 3 to retrieve the best BLAST hits in the *T. aestivum* genomes listed in Table 4-1 (http://www.10wheatgenomes.com/). As mentioned in the Chapter 1, these genomes have been assembled following a pipeline comparable to the one that produced Chinese Spring assembly RefSeqv1.0. These nine

assemblies are thus chromosome-scale assemblies. Given that these assemblies were released in 2019, we did not have access to them at the time we designed the *Yr7* and *Yr5* specific markers.

We used the same parameters as in Chapter 3 to identify potential *Yr7* and *Yr5* alleles. We retained as potential alleles BLAST hits that aligned across 99 % of the query length and shared at least 95 % identity with the query.

Table 4-1. Summary of the nine chromosome-scale wheat assemblies used in this chapter (http://www.10wheatgenomes.com/progress/)

Cultivar	Name used in this Chapter	Region
Julius	Julius	Germany
Jagger	Jagger	USA
Norin61	Norin61	Japan
CDC Landmark	Landmark	Canada
CDC Stanley	Stanley	Canada
Arina*LrFor*	Arina	Switzerland
Mace	Mace	Australia
SY-Mattis	SY-Mattis	France
Lancer	Lander	Australia

4.2.2. Defining the *Yr7*, *Yr5* and *YrSP* syntenic region in wheat and related species

4.2.2.1. Definition of syntenic regions across wheat pangenome

We used NLR-Annotator[195] to identify putative NLR loci on RefSeq v1.0 chromosome 2B and identified the best BLAST hits to *Yr7* and *Yr5* on RefSeq v1.0. Additional BED-NLRs and canonical NLRs were annotated in close physical proximity to these best BLAST hits. To define the syntenic interval encompassing the NLR cluster we selected ten non-NLR genes located both distal and proximal to the region, and identified

homologs in the nine assemblies described in Table 4-1. The non-NLR genes flanking RefSeq NLRs close the *Yr7* and *Yr5* best BLAST hits are listed in Appendix 8-15.

Definition of the gene content of the syntenic regions in the wheat genomes:

We extracted the full genome sequence starting from the most distal syntenic non-NLR gene distal to the *Yr7/5* locus and ending at the most proximal syntenic non-NLR gene proximal to the *Yr7/5* locus in all nine assemblies. There is no gene annotation currently available for the nine newly sequenced and assembled wheat genomes (Table 4-1). However, the Plant Genome and Systems Biology (PGSB, http://pgsb.helmholtz-muenchen.de) shared projections of the RefSeqv1.1 (Chinese Spring) gene models onto these nine assemblies (Manuel Spannagl, personal communication). These projections combine both sequence similarity and synteny information (surrounding genes) to assign a given RefSeqv1.1 gene model to a specific region on the nine genomes. We used this information to identify all projected gene models located in the *Yr7/5* syntenic interval in the nine genome assemblies.

Definition of the NLR content of the syntenic regions in the wheat genomes

Presence/absence variation across varieties of the same species is a known feature of NLR genes[252]. Given that it is likely that not all Chinese Spring NLRs will be present in these assemblies, we thus cannot rely alone on gene projections to define the NLR locus in the nine assemblies. In addition, these assemblies could contain NLRs that are too different from Chinese Spring to be part of the gene model projections.

Similar to what we did on RefSeqv1.0, we used NLR-Annotator to identify potential NLR loci in each of the syntenic regions derived from the nine genome assemblies. We looked for any of these loci overlapping with a projected gene model and kept the gene

structure derived from the gene model when possible. Some translated proteins derived from these projections contained premature termination codons, suggesting that there might be differences between the Chinese Spring gene structure and their best hits in the nine assemblies.

When no gene model was available, we carried out a 6-frames translation (https://www.ebi.ac.uk/Tools/st/emboss_transeq/) of the extended (+/- 1,000 bp) NLR-Annotator loci. We subsequently used hmmscan from HMMER v3.1[253] to compare these sequences with the Pfam database (ftp://ftp.ebi.ac.uk/pub/databases/Pfam) and identify additional domains, such as BED domains, in the gene models and the translated sequences. We applied a cut-off of 0.01 on i-evalue to filter out any irrelevant identified domains This allowed us to determine which NLR proteins were likely to BED-NLRs and which were canonical NLRs.

4.2.2.2. Definition of syntenic regions across grass genomes

We used our syntenic non-NLR gene set from Chinese Spring to define the *Yr7/5* region in barley, *Brachypodium*, and rice in *EnsemblPlants* (https://plants.ensembl.org/). We used different percentage identity cut-offs for each species based on how phylogenetically related to wheat they are (> 92 % for barley, > 84 % for *Brachypodium*, and > 76% for rice) and determined the syntenic region when at least three consecutive orthologues were found. A similar approach was conducted for *Triticum turgidum* ssp. *dicoccoides* and *Ae. tauschii* (Appendix 8-15). All investigated genomes are listed in Table 4-2.

Table 4-2. Summary of genome assemblies used to identify the *Yr* syntenic region in wheat related species.

Specie	Cultivar/group	Source	Link/ref
Triticum aestivum	Chinese Spring	IWGSC	https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies
Triticum turgidum	Zavitan	WEWseq	Avni et al. 2017[254]
Aegilops tauschii	AL8/78	UC Davis	Luo et al. 2017[255]
Oryza sativa	japonica	Ensembl / RAP-DB	http://plants.ensembl.org/Oryza_sativa/Info/Index
Brachypodium distachyon		Ensembl / Brachypodium.org	http://plants.ensembl.org/Brachypodium_distachyon/Info/Index
Hordeum vulgare	Morex	Ensembl / IBSC	http://plants.ensembl.org/Hordeum_vulgare/Info/Index

We used a similar approach to the one described in section 4.2.2.1 for the wheat pangenome to extract the *Yr7/Yr5* syntenic region in the grass genomes and annotate NLR loci with NLR-Annotator. We used previously defined gene models where possible, but also defined new gene models. These were further analysed through a BLASTx analysis to confirm the NLR domains (Appendix 8-15). The presence of BED domains in these newly annotated NLRs was confirmed with HMMERv3.1 and the Pfam database, as described in section 4.2.2.1.

4.2.3. Large-scale genomic comparison of the ten sequenced wheat genomes

4.2.3.1. Gene content comparison in the ten genomes

To determine the conservation of the wider *Yr7/Yr5* genomic region across different wheat varieties, we extended the syntenic region in Chinese Spring described above by 8 Mb at each end. We retrieved all annotated genes in this interval and performed a BLAST analysis to the other nine genomes (Table 4-1) to determine gene conservation across varieties. We used the following parameters to define the best BLAST hit for each gene:

- the target should cover at least 90 % of the query
- the target should be located on chromosome 2B and within the interval defined by the furthest best BLAST hits at both ends

To visualize the BLAST results, we generated a heatmap based on the percentage identity of each target with the corresponding query. We used ggplot2 R package[256] to draw the heatmap. We conducted this same approach with each of the nine genomes as the reference.

4.2.3.2. Yr7/5 region expanded alignments

To investigate whether the gene comparison above reflected how similar/distant varieties were beyond the immediate coding sequences, we performed pairwise alignments of the chromosome-scale assemblies between varieties showing high conservation at the gene level. We used the nucmer program of MUMmer v3.0[257] to perform the alignments and mummerplot to generate a gnuplot script to draw the corresponding alignment graphs. We did not filter on conservation before drawing the plot, as we wanted to see how similar/different the two varieties were in this region. We used dnadiff function of MUMmer to extract the alignment statistics for each alignment and determine the number of SNPs and associated SNP density within and outside the *Yr* locus. We identified repeat and low complexity regions with RepeatMasker v4.9.19 (www.repeatmasker.org) and the TREP database v20 (http://botserv2.uzh.ch/kelldata/trep-db/downloadFiles.html).

4.2.4. Phylogenetic analysis of the NLRs located in the *Yr* locus across grass species

We defined the *Yr7/Yr5* syntenic region in wheat and related species listed on Table 4-2 in section 4.2.2.2. We extracted all NB-ARC domains from the predicted proteins corresponding to the NLR loci and aligned them with MAFFT[258] using default parameters (v7.305). We verified and manually curated the alignment with Jalview[212] (v2.10.1). We used Gblocks[259] (v0.91b) with the following parameters: Minimum Number Of Sequences For A Conserved Position: 9; Minimum Number Of Sequences For A Flanking Position: 14; Maximum Number Of Contiguous Non-conserved Positions: 8; Minimum Length Of A Block: 10; Allowed Gap Positions: None; Use

Similarity Matrices: Yes; to eliminate poorly aligned positions. This resulted in 36% of the original 156 positions being taken forward for the phylogeny. We built a Maximum Likelihood tree with the RAxML[260] program and the following parameters: raxmlHPC -f a -x 12345 -p 12345 -N 1000 -m PROTCATJTT -s <input_alignment.fasta> (MPI version v8.2.10). The best scoring tree with associated bootstrap values was visualised and mid-rooted with Dendroscope[261] (v3.5.9).

4.2.5. Identifying BED-NLRs and BED-proteins in plant genomes

We downloaded 90 plant proteomes from Phytozome v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html) and *EnsemblPlants* (https://plants.ensembl.org/index.html) (Appendix 8-17) and identified complete **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs (BUSCO genes) with the BUSCO program[262] (v3). Given that we investigated proteomes from all plant kingdom, we performed two BUSCO analyses: one with the Viridiplanteae set[263], which comprises 430 orthologs, and one with the Embryophytes set[263], which comprises 1,440 orthologs. We filtered-out any proteome displaying less than 90 % of complete orthologs from the Viridiplanteae set and any Embryophyte proteome displaying less than 90 % of complete orthologs from the Embryophyte set. Our final set contained 68 proteomes (69 with RefSeqv1.0, Appendix 8-17).

For these 69 proteomes, we identified proteins carrying a BED domain with HMMER (v3.1) and the Pfam database as described in section 4.2.2.1. We separated the set between NLR and non-NLRs based on the presence of the NB-ARC. BED domains were extracted from the corresponding protein sequences based on the HMMER output. We retained a total of 20 proteomes containing both BED-NLRs and other BED-containing proteins for the Neighbour-net analyses.

4.2.6. Neighbour-net analyses

We used the Neighbour-net method[264] implemented in SplitsTree4[265] (v4.16) to analyse the relationships between BED domains from NLR and non-NLR proteins in wheat (Figure 4-14). We first retrieved all BED-containing proteins from RefSeq v1.0 identified in section 4.2.5. We then aligned the BED domains with MAFFT[258] (v7.305) and used this to generate a neighbour network in SplitsTree4 based on the uncorrected P distance matrix. We carried out identical analyses in the 20 proteomes containing BED-NLRs and BED proteins (section 4.2.5). We grouped together species that were close phylogenetically to increase the power of the analysis. This allowed us to identify BED domains from BED-NLRs sharing sequence similarities with BED-domains from other proteins.

Additionally, we investigated whether a certain class of BED-protein would tend to cluster more with BED-NLRs based on BED domain similarity. We retrieved any additional domain identified in the HMMER analysis within proteins whose BED domains clustered with BED-NLRs (Appendix 8-18) and carried out an exact Fisher's test to determine whether the proportion of a given domain in BED-protein clustering with BED-NLRs was higher than the proportion of this domain in BED-proteins in general.

4.2.7. Re-analysis of transcriptomic data

We used RNA-Seq data previously published by Dobon and colleagues[197]. Briefly, two RNA-Seq time-courses were used based on samples taken from leaves at 0, 1, 2, 3, 5, 7, 9, and 11 days post inoculation (dpi) for the susceptible cultivar Vuka and 0, 1, 2, 3, and

5 dpi for the resistant AvocetS-Yr5[197]. We used normalised read counts (Transcript Per Million, TPM) from Ramirez-Gonzalez et al. 2018[221] to produce the heatmap shown in Figure 4-15 with the pheatmap R package (v1.0.8). Transcripts were clustered according to their expression profile as defined by a Euclidean distance matrix and hierarchical clustering. Transcripts were considered expressed if their average TPM was ≥ 0.5 TPM in at least one time point.

We used the DESeq2 R package[266] (v1.18.1) to conduct a differential expression analysis. We performed two comparisons: (1) likelihood ratio test to compare the full model ~ Cultivar + Time + Cultivar:Time to the reduced model ~ Cultivar + Time to identify genes that were differentially expressed between the two cultivars at a given time point after 0 dpi (workflow: https://www.bioconductor.org/help/workflows/rnaseqGene/); (2) investigation of both time courses in Vuka and AvocetS-Yr5 independently to generate all of the comparisons between 0 dpi and any given time point, following the standard DESeq2 pipeline. Genes were considered as differentially expressed genes if they showed an adjusted *p-value* < 0.05 and a log2 fold change of 2 or higher.

4.3. Results

4.3.1. Variation in the *Yr7*, *Yr5* and *YrSP* syntenic region across the wheat pangenome

4.3.1.1. Yr7 and Yr5 alleles in the pangenome

In Chapter 3, we identified five different alleles for *Yr5* in wheat assemblies that were available at that time. This includes the functional Spelt-Yr5 and Spaldings Prolific-YrSP which we cloned and confer resistance against *Pst* and three additional alleles, Claire-Yr5, Cadenza/Paragon/Robigus-Yr5 (referred to as Cadenza-Yr5) and Kronos/Svevo-Yr5 (referred to as Kronos-Yr5), which have not been functionally tested. We identified only one *Yr7* allele in these same assemblies. Later on, during the PhD, nine new genome assemblies became available (Chapter 1, Table 4-1) and we thus explored these to determine whether we could identify new alleles for *Yr7* and *Yr5*.

Yr7: We found one alternate *Yr7* allele in Landmark, Stanley and Mace sharing 99.98 % sequence identity with Cadenza-*Yr7* (Table 4-3). This polymorphism leads to a single amino-acid change in one of the manually annotated LRR repeat in Cadenza-*Yr7*.

Yr5: A BLAST hit for *Yr5* in the syntenic region was found in Lancer. However, this locus shares 95.09 % identity with Spelt-Yr5 so it is unclear whether it is a true allele or a distant homolog. We extended the BLAST search to the whole pangenome and identified two *Yr5* alleles: one in Julius and Jagger and another in Arina and SY-Mattis. Surprisingly, Julius/Jagger-Yr5 was identical to Claire-Yr5 and Arina/SY-Mattis-Yr5 was identical to Cadenza-Yr5, both described in Chapter 3. However,

these alleles were located on Chromosome 2D in Julius/Jagger and Arina/SY-Mattis and not on Chromosome 2B as we would expect for *Yr5* alleles.

Table 4-3. *In silico* allele mining for *Yr7* and *Yr5* in the ten chromosome-quality wheat assemblies.
Percentage identity of the identified alleles and matching colours illustrate identical haplotypes. Only hits with > 95 % identity to either *Yr7* or *Yr5* are reported

Genome (chromosome)	% identity to Spelt-Yr5 (DNA)	% identity to Cadenza-Yr7 (DNA)	Comment
Arina (2D)	99.31	-	Identical to Cadenza-Yr5
SY-Mattis (2D)	99.31	-	Identical to Cadenza-Yr5
Julius (2D)	99.75	-	Identical to Claire-Yr5
Jagger (2D)	99.75	-	Identical to Claire-Yr5
Lancer (2B)	95.09	-	-
Stanley (2B)	-	99.98	New *Yr7* allele-
Mace (2B)	-	99.98	New *Yr7* allele
Landmark (2B)	-	99.98	New *Yr7* allele
Chinese Spring	-	-	-
Norin61	-	-	-

This result illustrates how important it is to have access to sequence information from distant wheat varieties to design diagnostic markers. For example, our *Yr7* markers designed in Chapter 3 are not able to discriminate between Cadenza-Yr7 and Landmark/Mace/Stanley-Yr7 (referred to as Landmark-Yr7). However, is Landmark - Yr7 functionally different from Cadenza-Yr7? This is an important question to answer because if both alleles are functionally identical, then it is an advantage that the marker can select both alleles. Alternatively, if both alleles are different in terms of response to *Pst* (either resistant/susceptible or can identify different *Pst* isolates), then the marker would need to be adapted. Exploring the *Yr7* and *Yr5* alleles in the wheat pangenome is thus critical to determine if they are of potential interest in breeding and to define the best strategies for effective deployment.

4.3.1.2. Comparison of the Yr locus in ten sequenced wheat varieties

NLRs often organise into clusters in plant genomes[186]. In Chapter 3 we found that *Yr7*, *Yr5* and *YrSP* resistances were mediated by single dominant genes. However, most of the characterised NLR with integrated domains (NLR-IDs) function in pair with another NLR protein (discussed in section 4.1.1). In these cases, the NLR-ID is involved in pathogen recognition and this interaction allows activation of its partner, leading to defense signalling. These characterised paired NLRs are often in close physical proximity and in a head-to-head orientation in their genomic context. As discussed at the end of section 4.1.3, although we did not find evidence of an additional NLR involved in *Yr7*-mediated resistance, we cannot discard the hypothesis that a potential redundant helper might be required. Additionally, mutation independent to the *Yr5* locus were identified in other Lemhi-Yr5 mutants[174]. We thus hypothesized that exploring the close-proximity of *Yr7* and *Yr5* in varieties carrying these genes (or alleles) will enable us to determine whether a potential partner is present.

To address this, we defined the syntenic region around the *Yr7/Yr5* locus in RefSeqv1.0 and nine wheat cultivars that were newly sequenced and assembled during this thesis (ten genomes investigated in total). We will refer to this interval that includes the *Yr7/ Yr5* locus and the distal and proximal regions including non-NLR genes as the ***Yr* region** (described in section 4.2.2.1).

The syntenic region is highly variable in size across the sequenced cultivars (from 2.2 Mb in Julius and Lancer to 4.7 Mb in Jagger). NLR clusters are defined as an uninterrupted sequence of NLR loci in a region, whose size varies according to the gene density of the genome. Thus, we first needed to identify the number of annotated genes

in the region and determine which of them were NLRs to defined potential NLR clusters in the *Yr* regions across the ten wheat genomes.

Gene content

We obtained projections of the RefSeqv1.1 (Chinese Spring) gene models onto the nine newly sequenced and assembled wheat genomes (See Methods 4.2.2.1). These projections combine both sequence similarity and synteny information (surrounding genes) to assign RefSeqv1.1 gene models to a position on the nine genomes. We used these data to estimate the gene content in the *Yr* region across the different genomes (Table 4-4). We identified from 32 to 54 gene models in the *Yr* syntenic region in the ten genomes (Table 4-4). There was a correlation between the size of the region and the number of gene models (Table 4-4). For example, we found 32 gene models in Julius (2.2 Mb region) and 54 in Jagger (4.7 Mb region). However, it is important to bear in mind that because these gene models are Chinese Spring projections, varieties closer to Chinese Spring are likely to carry a higher percentage of projected gene models that more distant varieties. This will be further explored in the following section.

NLR content

The use of projections carries the consequence that genes absent from Chinese Spring will be missed from this analysis. Presence/absence variation (PAV) is significantly enriched in NLRs in *Arabidopsis thaliana*[252]. Providing the same occurs in wheat, we would thus likely miss NLR content across the ten genomes if we only consider the Chinese Spring projections. To address this, we used NLR-Annotator[195] to predict putative NLR loci in the regions (Table 4-4 and Figure 4-3).

We found between 13 (Julius and Lancer) and 18 (Landmark, Mace and Stanley) putative NLR loci in the regions across the ten genomes, with only 4 to 10 loci overlapping with Chinese Spring gene models (Table 4-4). This supports our hypothesis that NLR loci will be missed if considering only the gene projections. The highest number of NLR loci was not found in the largest regions and vice versa. Indeed, Jagger, Arina and Chinese Spring, which are among the largest *Yr* regions (4.5 to 4.7 Mb) have similar number of predicted NLR loci to Julius and Lancer which have the smallest *Yr* regions (2.2 Mb, Table 4-4). This shows that there is no relation between the size of the region and the number of NLR loci. The NLR density thus varies between varieties.

BED-NLR content

Given that *Yr7*, *Yr5* and *YrSP* encode BED-NLRs, we searched for conserved domains in the predicted proteins derived from the gene model projections, or the 6-frame translation of the NLR-Annotator loci (Appendix 8-10). We predicted between six (Jagger, Arina and Chinese Spring) and ten (Mace) BED-NLRs in the region. More BED-NLRs were found in genomes with the higher numbers for total NLRs (Table 4-4). BED-NLRs thus seem to be highly represented in this region, as exemplified by the fact that 6 of the 16 BED-NLRs found in Chinese Spring[88] are present in this interval.

Table 4-4. Gene content variation in the ten wheat genomes (http://www.10wheatgenomes.com), including NLRs and BED-NLRs.
Gene annotation corresponds to the projections of the RenSeqv1.1 annotation onto the nine genomes.

Genome	region size (Mb)	#annotated genes	#NLR loci	#NLR loci overlapping with gene model	#BED-NLRs	#BED-NLRs overlapping with gene models
Jagger	4.7	54	14	9	6	3
Arina	4.6	54	14	8	6	3
Norin61	4.5	46	17	8	7	3
Chinese-Spring	4.5	42	14	10	6	3
Landmark	3.6	40	18	7	9	2
Mace	3.2	39	18	5	10	2
Stanley	3.2	39	18	5	9	1
SY-Mattis	3.1	41	15	5	8	1
Julius	2.2	34	13	5	9	2
Lancer	2.2	32	13	4	9	2

Yr syntenic region architecture

Among the ten genomes we investigated, there were some strong similarities in the overall architecture of the *Yr* region and between NLR located in the region. We classified these into three groupings (Figure 4-3):

Group 1: Lancer and Julius

Group 2: Chinese Spring, Arina, Jagger and Norin61,

Group 3: Stanley, Mace, SY-Mattis and Landmark

Group 1: The synteny was highly conserved between Lancer and Julius, although there was one NLR present in Julius and absent in Lancer (nlr_13). Additionally, most of their NLR loci were 100% identical (black lines on Figure 4-3).

Group 2: We observed high conservation in both the sequence and the order of the NLR loci in Arina, Norin61 and Chinese Spring (Figure 4-3). The syntenic order was similar

in Jagger, but fewer NLRs were identical in sequence to those in Arina, Norin61 and Chinese Spring.

Group 3: The syntenic order of the NLRs was also highly conserved between Landmark and Mace (Figure 4-3). There were a few re-arrangements between these two varieties and SY-Mattis and Stanley, even though most of the NLRs are identical across these four cultivars. The three genomes where we found an additional *Yr7* allele (Mace, Landmark and Stanley) all belong to this group.

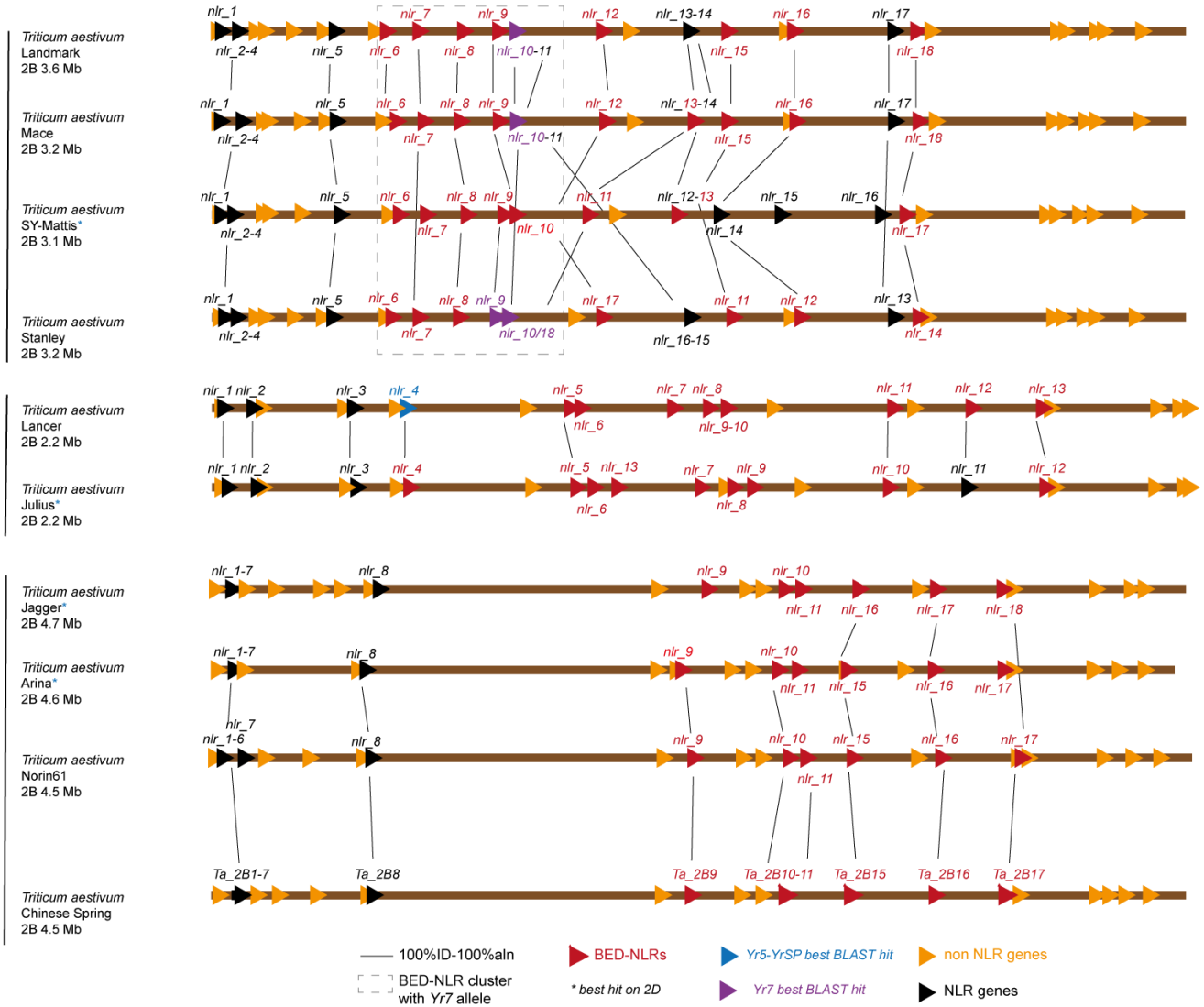


Figure 4-3. Schematics of the *Yr* syntenic region in ten sequenced wheat varieties.
Variety name and region size are shown on left. Triangles depict the genes annotated/projected (see section 4.2.2.1 for details) on each genome. Triangles show non-NLR genes (orange), canonical NLR genes (black) and BED-NLR genes (red). Orientation of the triangles does not reflect orientation in the genome. Blue and purple triangles depict BED-NLRs that are best BLAST hits for *Yr5* and *Yr7*, respectively. Black line linking triangles depict 100% identity across 100% of the sequence between the NLR genes (links for non-NLR genes not shown for clarity). Grey dashed box shows the BED-NLR cluster including the *Yr7* allele present in Landmark, Stanley and Mace.

Focus on the NLR cluster containing the *Yr7* allele in Landmark, Mace and Stanley (Group 3)

We identified small regions where only successive NLR loci uninterrupted by other genes were found (Figure 4-3). For example, nlr_6 to nlr_11 in Landmark, Mace, SY-Mattis and Stanley are located in a 430 kb-long region in Landmark and there is no other annotated gene between these BED-NLRs. Interestingly this potential cluster contains the *Yr7* allele (nlr_10) in Landmark, Mace and Stanley (Figure 4-3, Figure 4-4). nlr_10 is very close to the next NLR in Landmark/Mace (nlr_11, ~ 4.2 kb apart), which is very similar to what was observed for paired NLRs (section 4.1.1). We observed a similar arrangement in Stanley, although the distance between the two loci was longer (nlr_10/nlr_18, ~ 10 kb apart) (Figure 4-4). There was no locus as close to *Yr7* as nlr_10 (Landmark, Mace and Stanley) in SY-Mattis. Indeed, nlr_10 is SY-Mattis was more distant, although we found evidence of a BED domain in this locus as well (Figure 4-4).

We identified only NB-ARC and LRR regions in the translated sequence of nlr_11 in Mace and Landmark. However, we reported a BED domain in nlr_18 locus in Stanley (Appendix 8-10, Figure 4-4). Although, nlr_10 and nlr_11/nlr_18 are not in head-to-head orientation, as observed for paired NLRs, it would be interesting to determine whether nlr_11 encodes a functional protein (no projected gene model was overlapping with nlr_11/nlr_18). Additionally, nlr_10 and nlr_11 in Landmark and Mace share only ~79.3 % identity, whereas nlr_10 and nlr_18 in Stanley share 95.7 % identity.

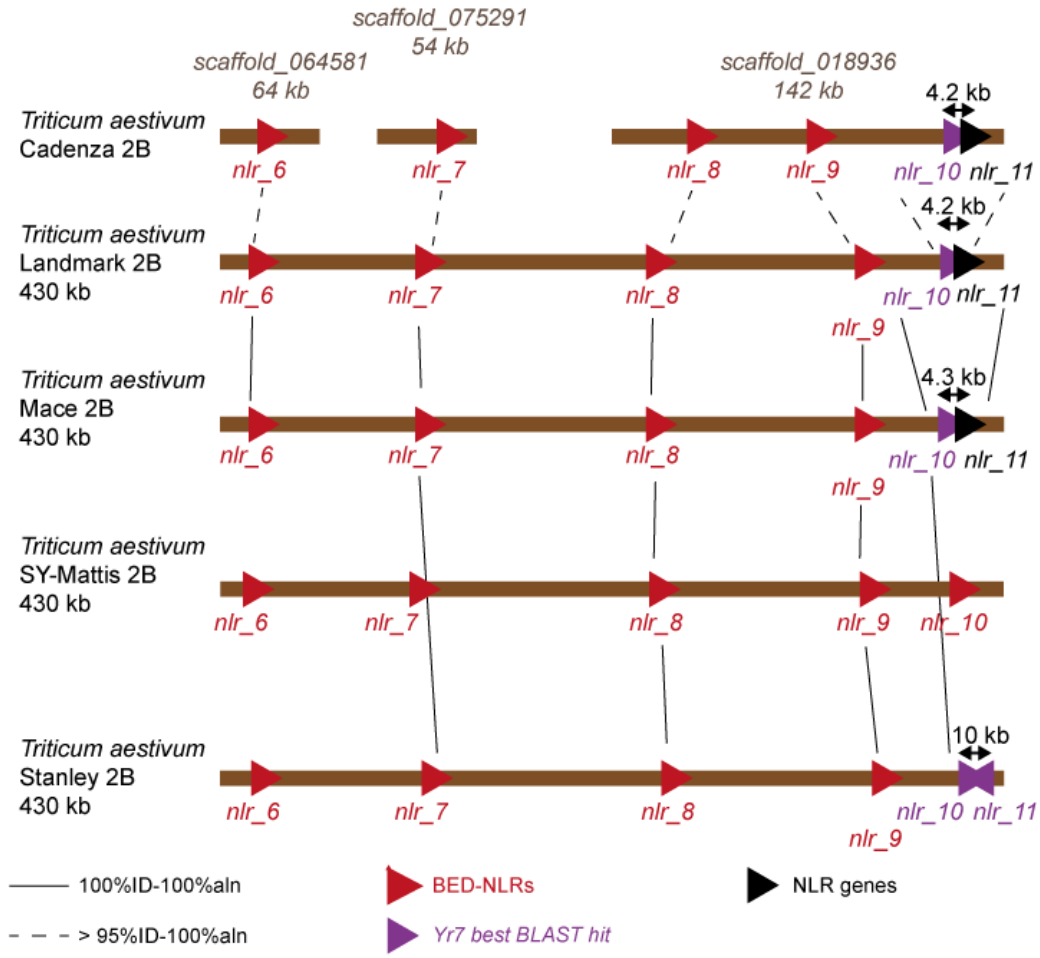


Figure 4-4. Close-up of BED-NLR cluster including the Yr7 allele in Landmark, Mace, and Stanley.

Variety name and region size are shown on left. Triangles show canonical NLR genes (black) and BED-NLR genes (red). Orientation of the red triangles does not reflect orientation in the genome. Purple triangles depict BED-NLRs that are best BLAST hits for *Yr7*. Black line linking triangles depict 100% identity across 100% of the sequence between the NLR genes and black dashed line > 95 % identity between BED-NLR from Landmark and BED-NLR from Cadenza. Distance between *Yr7* allele and the closest NLR is shown with a double-headed black arrow.

We did a BLAST search with nlr_11 Landmark in Cadenza scaffolds and identified a similar NLR in Cadenza (98.4 % sequence identity), located ~ 4.2 kb from *Yr7* and sharing ~ 79.7 % identity to *Yr7*. However, there were 'Ns' is the Cadenza sequence. Fortunately, there was a contig in the Cadenza RenSeq assembly that was 100 % identical to this sequence, excluding the Ns. We thus used this RenSeq contig to map RNA-Seq data from Cadenza to define exon/intron boundaries and predict the corresponding protein as described in Chapter 3. However, despite observing RNA-Seq reads mapping to the locus, the putative exons included premature termination codons so we could not predict a full-length protein (data not shown). Further investigation will thus be required to determine whether this potential pseudogene is functional.

Additionally, we identified nlr_6 to nlr_11 from Landmark in the Cadenza genome (> 95 % identity across 100 % of each locus) (Figure 4-4). The BED-NLRs were located in three different scaffolds, with scaffold_018939 containing nlr_8 to nlr_11 (142 kb long). The order is conserved in Cadenza, although the contig is slightly shorter than in Landmark (~ 184 kb from nlr_8 to nlr_11 in Landmark). We would have to anchor the whole region in Cadenza to determine whether nlr_6 and nlr_7 are in the same syntenic order than in Landmark. However, we decided to focus on the BED-NLRs in close proximity to the *Yr*7 allele here.

From this analysis we conclude that the *Yr7* allele in Landmark, Mace and Stanley is part of an NLR cluster on Chromosome 2B. This cluster is 430 kb long and contains six NLRs and no other predicted genes. However, as gene models in these varieties are projections from RefSeqv1.1 annotation, we cannot exclude that they could contain genes that are not present or are very different from Chinese Spring. We could confirm that the NLR in close proximity (~ 4kb) to *Yr7* allele in Mace and Landmark also exists in Cadenza.

Given that nlr_18 in Stanley is further apart from the *Yr7* allele and does not share similarities with nlr_11 in Cadenza, Mace and Landmark, the architecture of this cluster is slightly different in Stanley. It would thus be interesting to test whether Stanley still displays a *Yr7*-like phenotype to determine whether nlr_11 in Cadenza/Landmark/Mace is important for the expression of *Yr7* resistance.

4.3.2. Comparison of the expanded *Yr* locus in ten sequenced wheat genomes

In the section above, we observed variation in the *Yr* region across the wheat pangenome, although we could define three sub-groups across these sequenced genomes based on some degree of gene conservation (Figure 4-3). Is this conservation due to an overall similarity between the varieties in a sub-group? Or do these NLR-enriched regions behave differently from the surrounding genomic context in terms of conservation between cultivars? This led us to ask the following question, is the variation in NLR-enriched regions within a subgroup similar to that in the surrounding genomic context? More specifically, are cultivars belonging to the same sub-group identical by descent? We hypothesized that it is the case. In this Chapter, we chose to focus on the *Yr* locus to make the first observations that will allow us to fine-tune this hypothesis in future investigations.

We investigated the expanded syntenic region (see section 4.2.3) surrounding the *Yr* locus to determine whether varieties that are similar in this locus share identity outside of it and vice-versa. In this section, we will refer to this as the **extended *Yr* region**. For this analysis we did not consider NLRs only but the whole annotated and projected gene annotation (BLAST analysis) and the intergenic regions (genome alignments).

4.3.2.1. Comparison of gene content in the expanded Yr region between Chinese Spring and nine wheat varieties

We first used Chinese Spring as a reference to determine the similarities between the wheat reference genome and the other sequenced varieties. We extended the boundaries of the *Yr* regions defined in the section above by 8 Mb on each side, extracted all gene models that were located in this ~ 20 Mb interval (200 gene models plus 17 NLRs in the *Yr* region defined in Figure 4-3) and performed BLAST analysis with these genes against the nine other genomes. We filtered out all hits that were not overlapping at least 90 % of the query and that were not located on chromosome 2B (Figure 4-5).

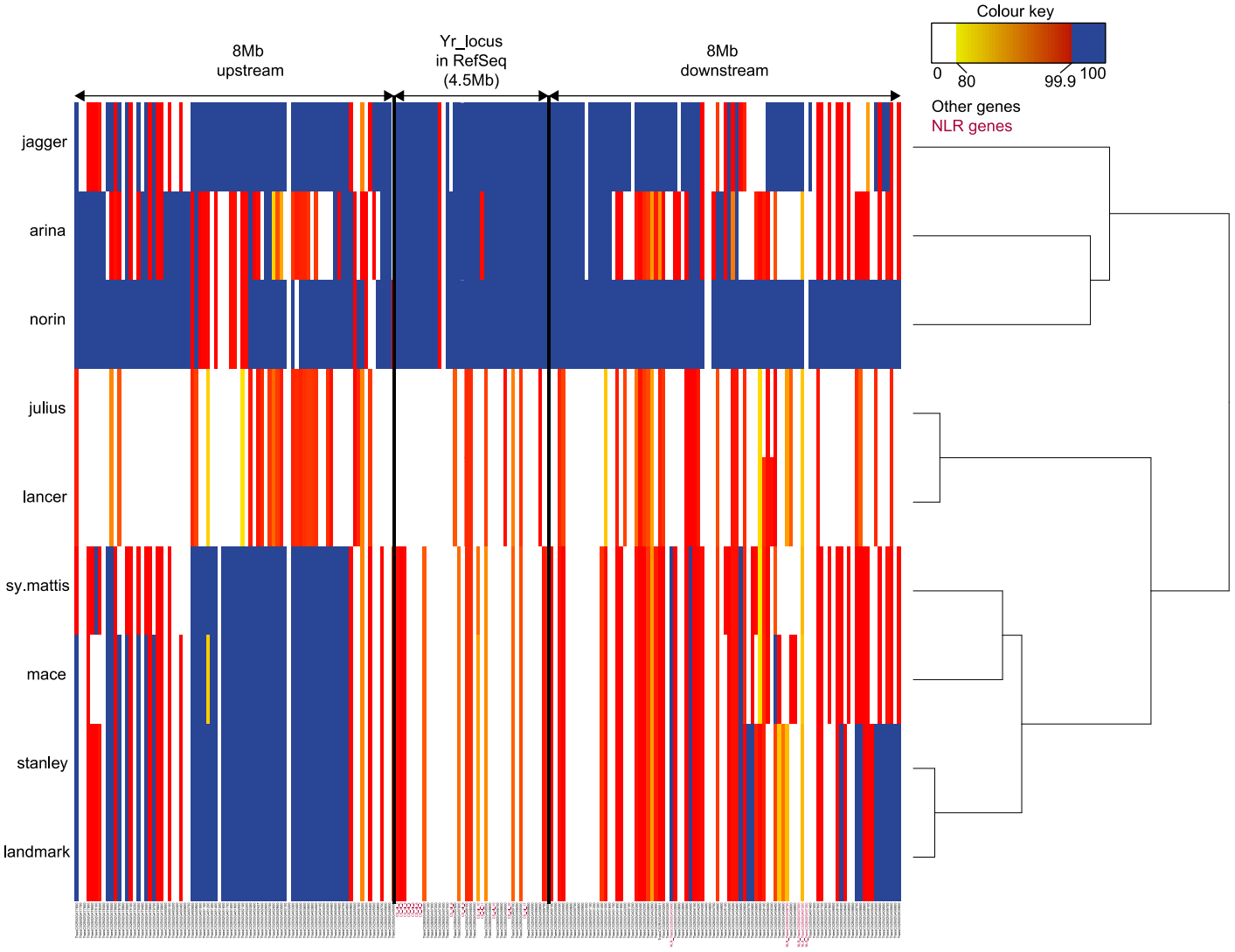


Figure 4-5. Heatmap of the BLAST analysis between Chinese Spring gene models in the expanded interval surrounding the *Yr* region and the nine other wheat genomes.
Only hits that overlapped at least 90 % of the query and located on chromosome 2B are displayed. The colour key ranges from white (no hit; < 80 % identity), yellow (close to 80 % identity), red (close to 99.9 % identity) to blue (strictly 100 % identity). Black arrows on the left show the boundaries of the region. Gene identifiers on the right indicate non-NLR genes (black) and NLR genes (dark red). Results were clustered according to the hierarchical clustering methods implemented in the heatmap 2 function of R (gplots package, v3.0.1.1.). Upstream refers to the proximal region (closer to the centromere) and downstream refers to the distal region (distant from the centromere).

Overall the results agree to those observed in Figure 4-3. Jagger, Arina and Norin61 (Group 2) show more identical hits to Chinese Spring than the other genomes. Julius and Lancer (Group 1) were more distant to Chinese Spring, but displayed a nearly identical pattern when compared to each other. We could observe the same pattern for SY-Mattis, Mace, Stanley and Landmark (Group 3). These first results thus seem to support our hypothesis that varieties that are similar in the *Yr* region also tend to me more similar than others in the extended region, at least in the sequence similarity across genes.

We then addressed the question whether the observations derived from the BLAST analyses focused on gene models were sufficient to conclude on similarity between different genomes. To do this, we performed whole genome alignment of the extended *Yr* region between varieties that were hypothesized to be similar on Figure 4-3 and Figure 4-5.

4.3.2.2. Focused analysis on Arina, Jagger, Norin61 and Chinese Spring (Group 2)

We analysed Group 2 and generated a similar heatmap to Figure 4-5, except for the reference being Arina (Figure 4-6, Appendix 8-12). We changed the reference to validate whether Arina, Norin61, Jagger and Chinese are similar or if using Chinese Spring as a reference masked the differences between Arina, Norin61 and Jagger. We also added Cadenza because it carries *Yr7*. It is important to note, however, that the Cadenza assembly is not chromosome-assigned. Thus, it was not possible to select for BLAST hits exclusively on chromosome 2B and this was likely to generate 'noise' (incorrect assignments) on the heatmap. The patterns observed on the heatmap confirmed what we previously recorded in Figure 4-5 with Jagger, Norin61 and Chinese Spring being the closest varieties to Arina. The heatmap allowed us to define a more highly conserved

7.02 Mb region between Arina and Norin61, Jagger and Chinese Spring (dashed line on Figure 4-6), which includes the *Yr* locus.

To further determine whether the BLAST analysis focusing on gene sequence alone is a good indicator of the degree of conservation between genomes, we performed the whole genome alignment of the 20.6 Mb region shown in Figure 4-5. This would allow us to determine if the regions that were less conserved at the gene level were also less conserved at the whole genome level (Figure 4-6, right panel). We confirmed that the regions displayed as highly conserved at the gene level were also highly conserved in the whole genome alignment (Table 4-5, dashed black line and green line on Figure 4-6).

Table 4-5. Summary of the percentage identity within and outside the *Yr* locus between Arina and Chinese Spring, Jagger and Norin61.
Statistics were obtained with MUMmer v3.0[257].

	Length of *Yr* locus in reference	**%ID (upstream *Yr* locus)**	**%ID (*Yr* locus)**	**%ID (downstream *Yr* locus)**
Arina/Chinese-Spring	4.6	99.947	99.807	99.813
Arina/Jagger	4.6	98.902	99.814	99.788
Arina/Norin61	4.6	99.779	99.684	99.934

It is important to note that the proportion of repeated regions in the aligned region was very similar to what we observe in wheat in general (close to 80 %, Appendix 8-11). It is thus somewhat unexpected to observe such degree of conservation between two varieties given the high proportion of repeated elements in the analysed region. However, there was a large deletion (~5 Mb) in Norin61 when compared to Arina (Figure 4-6, right). It is not possible to identify such a deletion on the corresponding heatmap which is only based on genes, although we can observe a few missing genes on the heatmap in Norin61 (white colour, Figure 4-6, left). It is thus difficult to see these large deletions or re-arrangements with the gene-based BLAST analysis only.

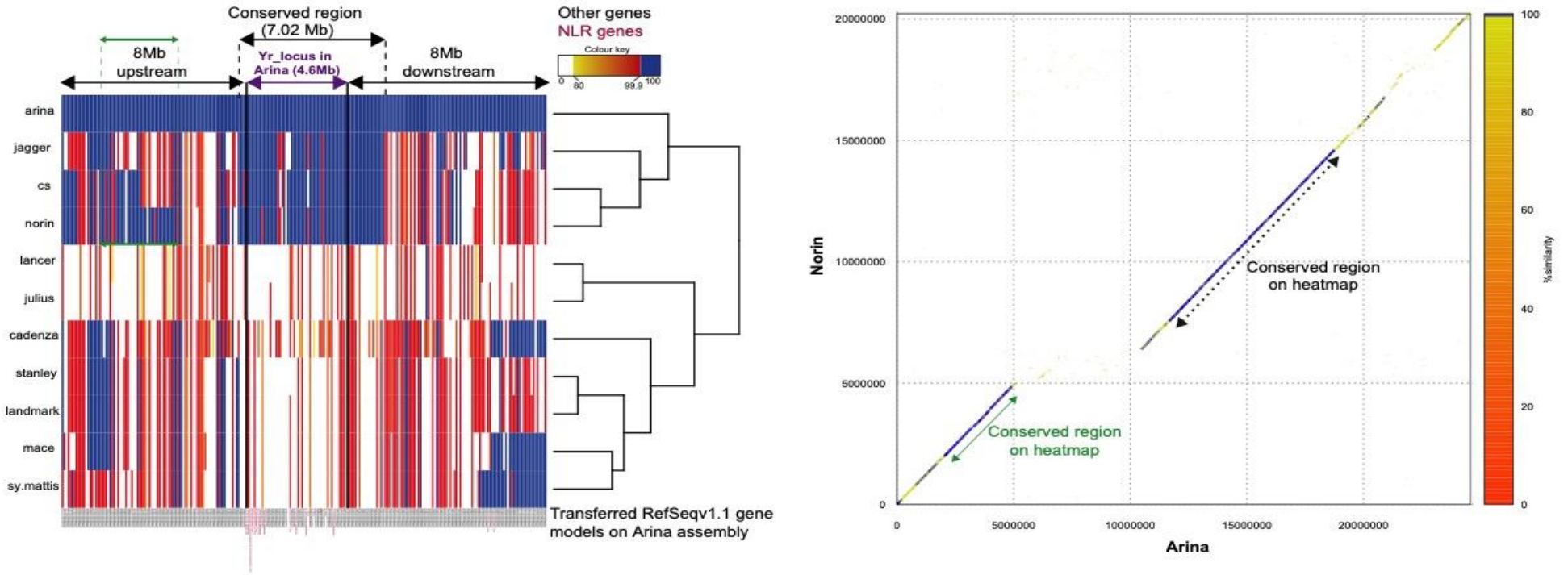


Figure 4-6. Heatmap illustrating the results of the BLAST analysis between Chinese Spring gene models and the nine wheat genomes + Cadenza (left) and alignment of a 21 Mb region in Arina and Norin61 (right).

Left: Only hits that overlapped > 90 % of the query and located on chromosome 2B are displayed. The colour key ranges from white (no hit; < 80 % identity), yellow (close to 80 % identity), red (close to 99.9 % identity) to blue (strictly 100 % identity). The black arrows at the top show the boundaries of the region with the location of the *Yr* region. The black dashed line shows the boundaries of the "Conserved region" between Arina and Norin61 (right). Gene identifiers are displayed at the bottom (non-NLR (black); NLR (dark red)). Results were clustered according to the hierarchical clustering methods implemented in the heatmap 2 function of R (gplots package, v3.0.1.1.). Upstream refers to the proximal region (closer to the centromere) and downstream refers to the distal region (distant from the centromere).

Right: alignment of the 20.4 Mb region showed on the heatmap (left) between Arina and Norin61 (see Appendix 8-12 for alignments of conserved region only between Arina and Chinese Spring/Jagger/Norin61). Blue colour shows region sharing a percentage identity higher than 99.9 %. Dashed line corresponds to the highly conserved region delimited with dashed lines on the heatmap and similarly for the region depicted with green arrows. The alignment was performed with MUMmer v3.0.

We then extracted the conserved region overlapping with the *Yr* region in each of these four varieties to perform pairwise whole genome alignment with Arina as the reference (Appendix 8-12). This close-up confirmed what we observed on Figure 4-6 regarding the high conservation within the region (99.7 % identity in the aligned region, Appendix 8-11), although some very localised sequences had much lower identity or did not even align. When overlaying the location of the NLR loci with the alignment (vertical black lines on Appendix 8-12), we observed that the region surrounding these loci was also conserved, which is consistent with the analysis shown on Figure 4-3.

Additionally, the SNP density in both the extended region and the *Yr* locus was similar in all comparisons (Table 4-6). Thus, in this particular group the variation in NLR-enriched regions is similar to that in the surrounding genomic context.

In Group 2, studying both NLR and non-NLR gene loci conservation across genomes was fairly consistent with the degree of conservation in whole genome alignments. However, large re-arrangements such as the large deletion in Norin61 were not immediately obvious on the heatmap. Thus, depending on the hypothesis, both analyses may be required.

Table 4-6. Comparison of number of SNPs and associated SNP density (#SNPs/Mb) between the whole extended *Yr* region and within the *Yr* locus defined in section 4.2.2.1 in Arina, Chinese Spring, Norin61 and Jagger.
The last column represents the weighted number of SNPs if the SNP density in the *Yr* locus covered a region of the same size of the extended region (7.02 Mb in this case). Statistics were obtained with MUMmer v3.0[257]

	#SNPs incl. InDel (excluding Ns)	**SNP density (#SNP/Mb)**	**#SNPs incl. InDel (*Yr* locus)**	**SNP density *Yr* locus (#SNP/Mb)**	**Weighted #SNPs based on *Yr* locus SNP density (difference with observed #SNP)**
Arina/Chinese-Spring	2639	380	1712	372	2586 (-53)
Arina/Jagger	2760	397	1799	391	2718 (-42)
Arina/Norin	2883	414	1904	414	2886 (+ 3)

4.3.2.3. Focused analysis on Julius and Lancer (Group 1)

Julius and Lancer were highly similar in the *Yr* region (Figure 4-3) and shared a nearly identical pattern of gene sequence similarity against Chinese Spring (Figure 4-5). We thus hypothesized that these two varieties were identical by descent in the observed region. To test this, we aligned the whole region surrounding the *Yr* locus in Julius and Lancer (Figure 4-7, Appendix 8-13).

Similarly, to what we observed on the heatmap (Appendix 8-13), the extended *Yr* region in Julius and Lancer (12.6 Mb) was highly similar at the genomic level (99.94 %, Table 4-7, Appendix 8-11). This included 79 % of repetitive regions (Appendix 8-11). The very low SNP density recorded in Appendix 8-11 is thus consistent with this observation (Table 4-8, 21 SNPs/Mb). However, the SNP density recorded in the 2.2 Mb corresponding to the *Yr* region was higher (62 SNPs/Mb, Appendix 8-11). The *Yr* region thus seemed conserved overall but carried noticeable difference between Julius and Lancer. This is different from what we observed for Arina and Jagger/Chinese Spring/Norin61 where a similar SNP density was observed in the *Yr* region compared to the average across the whole alignment (Appendix 8-11).

Table 4-7. Summary of the percentage identity within and outside the *Yr* locus between Julius and Jagger.
Statistics were obtained with MUMmer v3.0[257]

	Length of *Yr* locus in reference	**%ID (upstream *Yr* locus)**	**%ID (*Yr* locus)**	**%ID (downstream *Yr* locus)**
Julius/Lancer	2.2	99.959	99.902	99.954

Table 4-8. Comparison of number of SNPs and associated SNP density (#SNPs/Mb) between the whole extended *Yr* region and within the *Yr* locus defined in section 4.2.2.1 in Julius and Lancer.
The last column represents the weighted number of SNPs if the SNP density in the *Yr* locus covered a region of the same size of the extended region (12.85 Mb in this case). Statistics were obtained with MUMmer v3.0[257]

	#SNPs incl InDel (excluding Ns)	**SNP density (#SNP/Mb)**	**#SNPs incl InDel (*Yr* locus)**	**SNP density *Yr* locus (#SNP/Mb)**	**Weighted #SNPs based on *Yr* locus SNP density (difference with observed #SNP)**
Julius/Lancer	266	21	137	62	784 (+518)

Within the *Yr* region, NLR positions co-localised with small structural re-arrangements, whereas the rest of the alignment was highly contiguous between Julius and Lancer. However, when investigating the ends of such re-arrangements, they occurred in regions harbouring several 'Ns' in one assembly or the other. We thus cannot determine whether such arrangements are real or due to mis-orientation/positioning of the contig in a given assembly. This illustrates how difficult it can be to resolve genomic region carrying NLRs organised in clusters.

Overall and similarly to the analysis on Arina, the BLAST analysis based on gene loci reflects the degree of conservation between Julius and Lancer in whole genome alignments. Furthermore, Julius and Lancer seem to be identical by descent, although they exhibit some variation within the *Yr* locus.

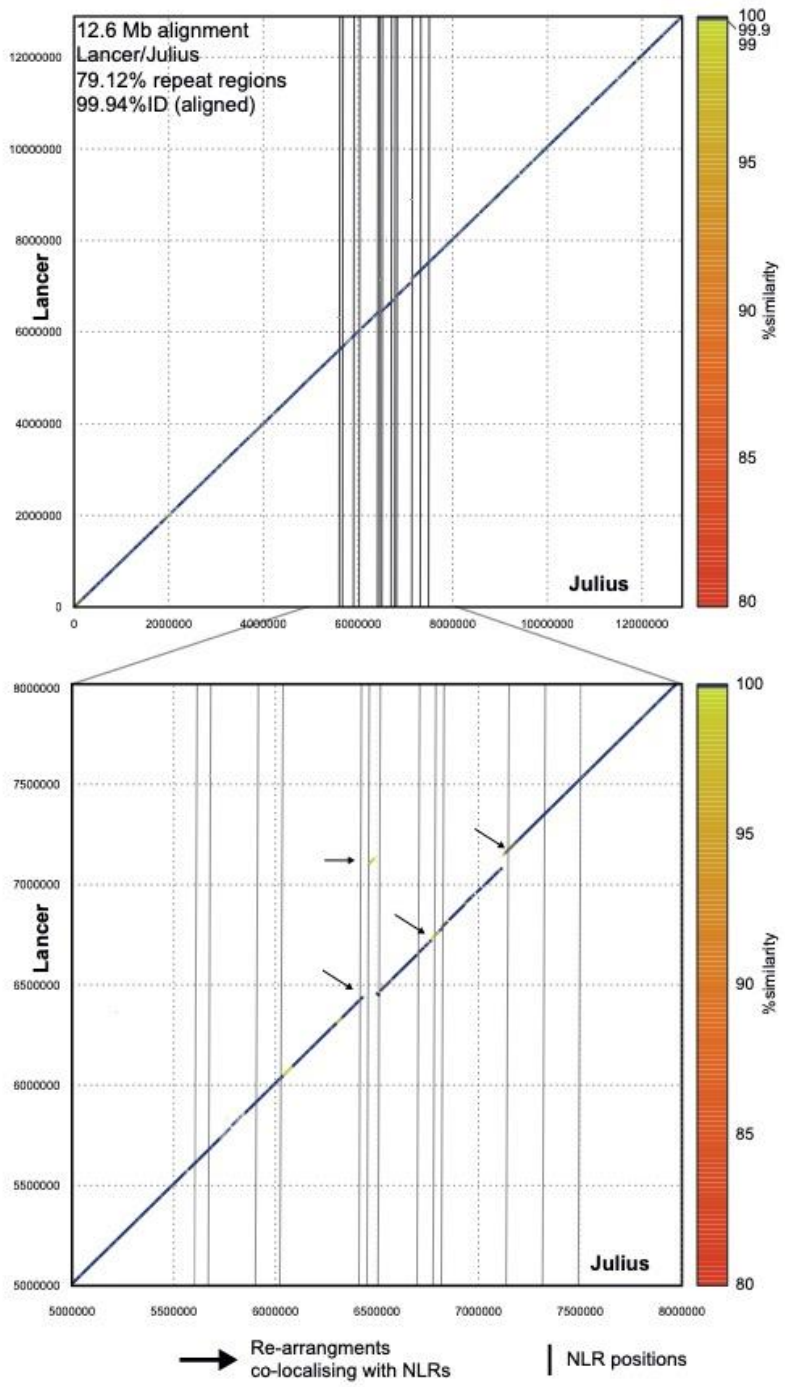


Figure 4-7. Alignment between Julius and Lancer in the region surrounding the *Yr* regions.

Top: alignment of the whole region (12.6 Mb) with vertical bars depicting NLR loci locations. Blue colour in the alignment refers to regions showing more than 99.9 % identity between the two genomes.

Bottom: close-up in the 2.2 Mb region encompassing the BED-NLR region in Figure 4-3. Vertical bars show NLR loci locations and arrows point to re-arrangements co-localising with NLR positions.

4.3.2.4. Focused analysis on Landmark, Mace, Stanley and SY-Mattis (Group 3)

We previously showed that Landmark, Mace, Stanley and SY-Mattis were broadly similar in the *Yr* region (Figure 4-3). To further assess this across the wider region, we carried out the genomic alignment analysis as above. The heatmap on Figure 4-8 shows that Landmark is highly similar to Stanley across the 18.6 Mb interval encompassing the *Yr* region. Indeed, there were only five different gene loci across the 18.6 Mb analysed, from which two were in the *Yr* region. Mace was identical to Landmark in the *Yr* region, as shown on Figure 4-3, but showed divergence in both the proximal and distal sections of the investigated region. SY-Mattis was the most distant variety, as we previously showed on Figure 4-3. There was a high degree of variation between Landmark/Mace/Stanley (alternate *Yr7* allele) and Cadenza (cloned *Yr7* allele) in the 3.6 Mb *Yr* locus. Indeed, it seemed that many loci were missing in the Cadenza assembly. However, it is important to note that *Yr7* and the closest NLR were both carrying 'Ns' in their corresponding loci and such hits would have been filtered out in this analysis. Overall all varieties were highly similar based on the whole aligned region (Table 4-9).

Table 4-9. Summary of the percentage identity within and outside the *Yr* locus between Landmark, SY-Mattis, Mace and Stanley.
Statistics were obtained with MUMmer v3.0[257]

	Length of *Yr* locus in reference	**%ID (upstream *Yr* locus)**	**%ID (*Yr* locus)**	**%ID (downstream *Yr* locus)**
Landmark/SY-mattis	3.6	99.643	99.925	99.948
Landmark/Mace	3.6	95.515	99.972	99.977
Landmark/Stanley	3.6	99.609	99.329	99.958

Given that Landmark and SY-Mattis seemed to be the most distant varieties from this group according to the heatmap in Figure 4-8, we investigated their conservation at the genomic level to determine whether the heatmap was a good representation of the variation between these two varieties. The genome alignment between Landmark and SY-Mattis on Figure 4-8 was overall consistent with what we observed on the heatmap.

There was a large conserved region between the two varieties that corresponded with the gene loci in blue on the heatmap. Both ends of the region were more variable in the heatmap and the alignment. However, it seems that at the start of the region on the heatmap (green box on Figure 4-8) there are numerous fairly conserved gene loci, whereas the beginning of the alignment does not show any collinearity on the dot plot. Thus, in this case, the heatmap was not very representative of the actual variation at the genomic level between SY-Mattis and Landmark. We observed the same between Mace and Landmark (Appendix 8-14), where all genes were 100 % identical between the two varieties at the boundary of the aligned region (heatmap, Appendix 8-14) but the corresponding region did not show overall alignment (top right, Appendix 8-14). It is, however, important to note that apart from being filtered for being located on chromosome 2B, there is no synteny restriction on the BLAST hits. This means that the Landmark gene model could hit SY-Mattis chromosome 2B at another location, for example, beyond the region being investigated here and still show a high percentage identity. This explains why both BLAST hits and alignment information are important.

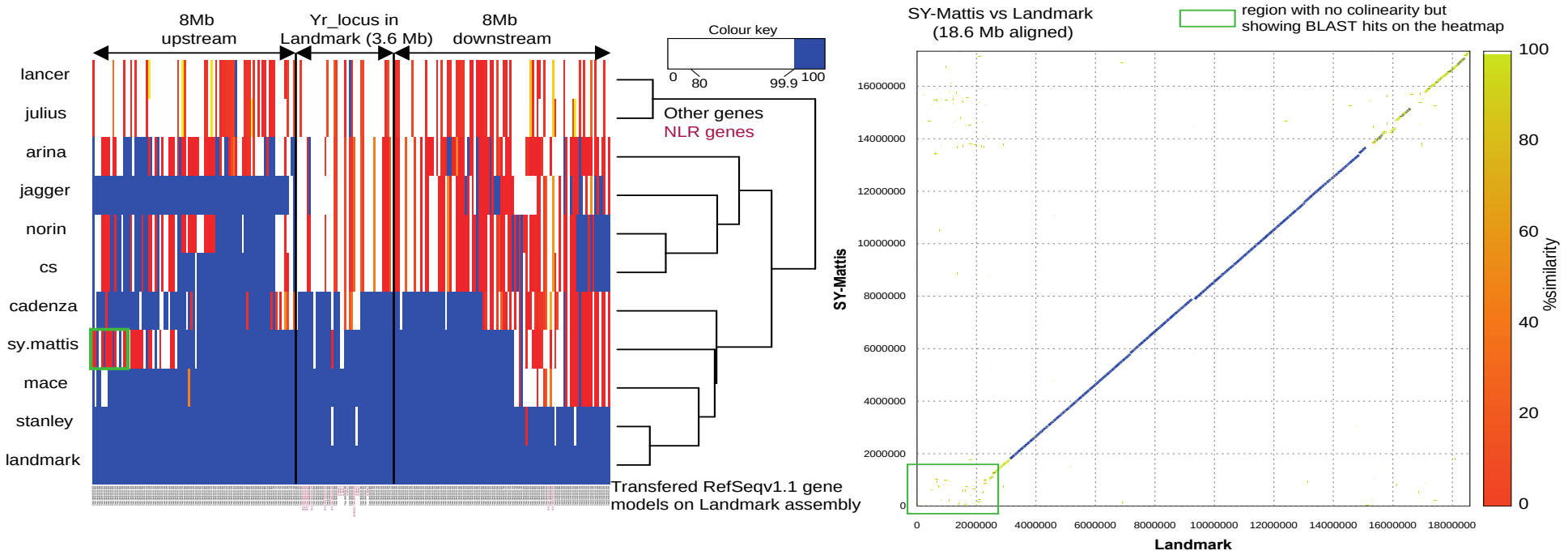


Figure 4-8. Heatmap illustrating the results of the BLAST analysis between Landmark gene models located in the expanded interval surrounding the *Yr* region and the nine other wheat genomes + Cadenza (left) and alignment of the 18.6 Mb region in Landmark and Norin61 (right).
Left: Only hits that overlapped at least 90 % of the query and located on Chromosome 2B are displayed. The colour key ranges from white (no hit to < 80 % identity hits), yellow (close to 80 % identity hits), red (close to 99.9 % identity hits) to blue (strictly 100 % identity hits). The black arrows at the top show the boundaries of the region with the location of the *Yr* region. Gene identifiers are displayed on at the bottom with black indicating a non-NLR locus and dark red a NLR locus. Results were clustered according to the hierarchical clustering methods implemented in the heatmap 2 function of R (gplots package, v3.0.1.1.). Green box corresponds to the gene loci falling into the corresponding green box on the alignment plot. Upstream refers to the proximal region (closer to the centromere) and downstream refers to the distal region (distant from the centromere).
Right alignment of the whole 19.6 Mb region showed on the heatmap (left) between Landmark and SY-Mattis (see Appendix 8-14 for close-up alignments of conserved region only between Landmark and SY-Mattis/Mace /Stanley). Blue colour shows region sharing a percentage identity higher than 99.9 %. The alignment was performed with MUMmer v3.0[257]

We then investigated the alignment between Landmark and Stanley, which seemed to be highly similar from the heatmap (Figure 4-8). We aligned a 9.5 Mb region surrounding the *Yr* locus between these two varieties (Figure 4-9, Appendix 8-14). Interestingly, there was a large inversion between the two cultivars and both ends were very close to NLR loci (Figure 4-9). However, these ends corresponded to 'Ns' in Stanley. Despite the inversion being very large and overlapping several scaffolds, we cannot determine whether it is a real structural variation between Landmark and Stanley or an artefact due to an assembly error. Similarly to what we showed in Julius and Lancer (Group 2), this demonstrated the difficulty of resolving physical contiguity across NLR clusters.

It was striking how the SNP density was different in the *Yr* locus compared to the whole region in Landmark, Mace, Stanley and SY-Mattis (Table 4-10, Appendix 8-11). Indeed, SNP density was only 5 SNPs/Mb in the *Yr* locus (3.6 Mb) in Mace whereas it reached 406 SNPs/Mb on average across the 9 Mb aligned. We observed a similar trend in Stanley and SY-Mattis, although SY-Mattis did have more SNPs in the *Yr* locus, as expected from what we observed on Figure 4-3. Indeed, this variety also lacked the *Yr7* allele and its neighbouring canonical NLR as observed in Landmark, Mace and Cadenza (Figure 4-4). Overall this is a different trend to what we observed in Arina, Chinese Spring, Norin61 and Jagger, where SNP density both inside and outside the *Yr* locus were comparable. It was also different to Julius and Lancer, where the *Yr* locus was the most variable region (Appendix 8-11).

Table 4-10. Comparison of number of SNPs and associated SNP density (#SNPs/Mb) between the whole extended *Yr* region and within the *Yr* locus defined in section 4.2.2.1 in Landmark, SY-Mattis, Mace and Stanley.
The last column represents the weighted number of SNPs if the SNP density in the *Yr* locus covered a region of the same size of the extended region (9.7 Mb in this case). Statistics were obtained with MUMmer v3.0[257]

	#SNPs incl InDel (excluding Ns)	**SNP density (#SNP/Mb)**	**#SNPs incl InDel (*Yr* locus)**	**SNP density *Yr* locus (#SNP/Mb)**	**Weighted #SNPs based on *Yr* locus SNP density (difference with observed #SNP)**
Landmark/SY-mattis	745	78	172	48	456 (-289)
Landmark/Mace	3651	406	19	5	45 (-3606)
Landmark/Stanley	348	37	26	7	67 (-281)

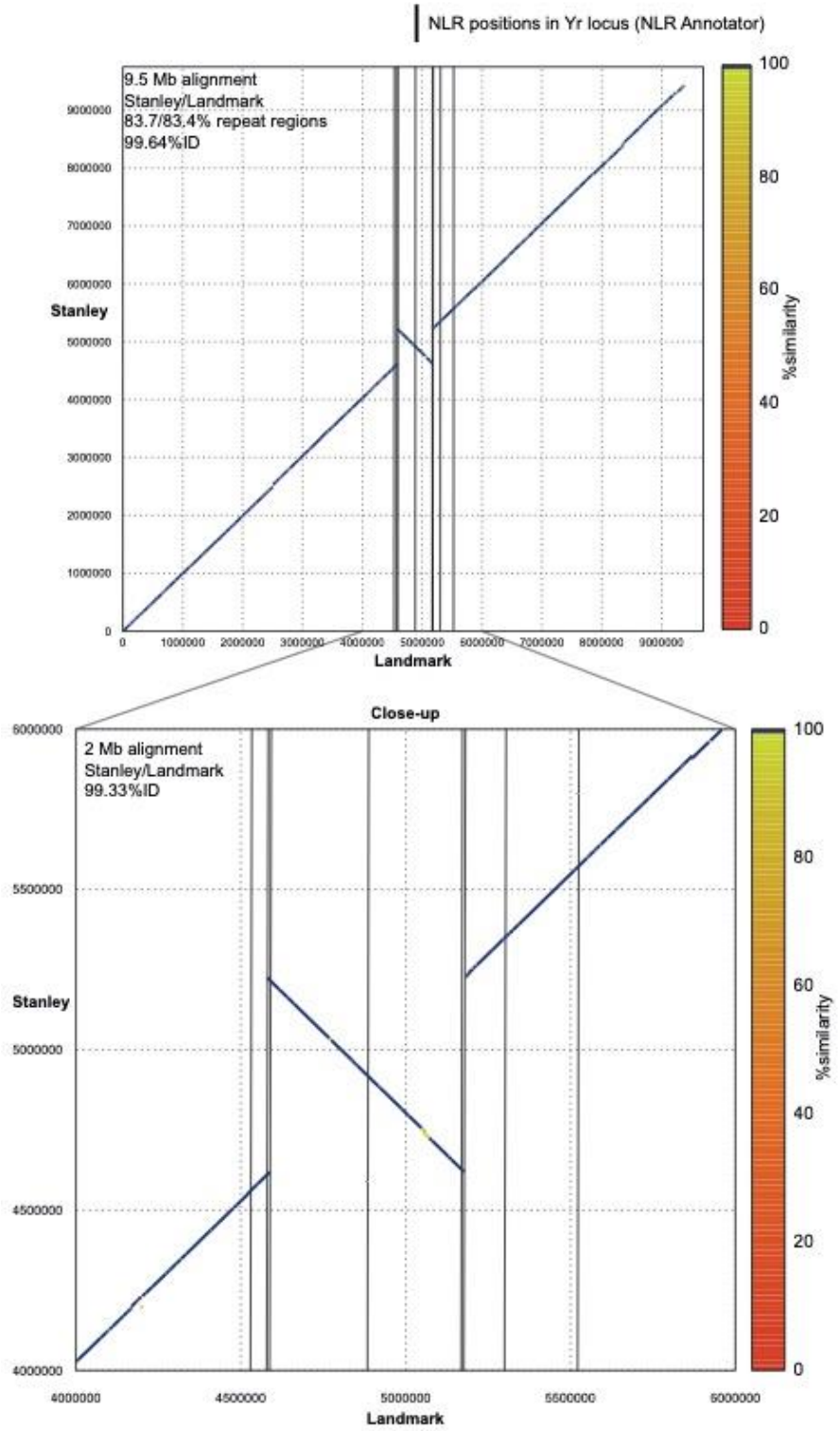


Figure 4-9. Alignment between Landmark and Stanley in the region surrounding the *Yr* regions.

Top: alignment of the whole region (9.5 Mb) with vertical bars depicting NLR loci locations. Blue colour in the alignment refers to regions showing more than 99.9 % identity between the two genomes.

Bottom: close-up in the 2 Mb region encompassing the BED-NLR region in Figure 4-3. Vertical bars show NLR loci locations

4.3.2.5. Summary

We analysed the conservation within and beyond the *Yr* locus in the three sub-groups we identified on Figure 4-3 to determine the degree of sequence conservation. Overall, we observed three different trends: no noticeable difference between within the *Yr* locus and outside (Arina/Jagger/Chinese Spring/Norin61), more variation within the *Yr* locus than outside (Julius/Lancer) or less variation within the *Yr* locus than outside (Landmark/Mace/Stanley/SY-Mattis).

Our working hypothesis was that the degree of conservation in the *Yr* region was similar to that of its surrounding genomic region. We showed in this section that it was not always the case. Investigating more NLR loci and their neighbouring regions across wheat varieties might help refining the hypothesis.

4.3.3. Analyses of the *Yr* locus in Chinese Spring (RefSeqv1.0) and related grass species

Analyses from the wheat pangenome enabled us to discover an additional allele for *Yr7* on chromosome 2B (Landmark, Mace and Stanley) and *Yr5* alleles identical to Claire-Yr5 (Julius, Jagger) and Cadenza-Yr5 (Arina, SY-Mattis) on chromosome 2D (Table 4-3). Additionally, we observed that the *Yr* locus was present in all investigated varieties, although varying noticeably between varieties from different sub-groups (Figure 4-3). Given such conservation in wheat, we asked the question whether the *Yr* locus, including BED-NLRs, would be present in related grass species. Identifying such conserved structure across related grass species would suggest that this architecture has been selected across the grass divergence.

4.3.3.1. *Definition of the Yr locus in wheat and related grass species*

We defined the *Yr* locus syntenic region in wheat and related species as described in section 4.2.2.2 and Appendix 8-15. We included all three homoeologous chromosomes from *Triticum aestivum* cv. Chinese Spring, the two homoeologous chromosomes from *Triticum turgidum* ssp. *dicoccoides* cv. Zavitan (wild emmer, durum wheat ancestor), the D genome progenitor of wheat *Aegilops tauschii*, *Hordeum vulgare*, *Brachypodium distachyon* and *Oryza sativa* japonica (Figure 4-10). The *Yr* locus was conserved between chromosome 2B of Chinese Spring and Zavitan. Additionally, we observed BED-NLRs in *Aegilops tauschii* and chromosome 2D of Chinese Spring, as well as on chromosome 2A of Zavitan and Chinese Spring. Moreover, such domain organisation was also observed in *B. distachyon* and *O. sativa* (Figure 4-10, Table 4-11). Additionally, there was an expansion of the number of BED-NLRs in wheat and wild emmer (Figure 4-10, Table 4-11) as compared to rice, *B. distachyon* and *Ae. tauschii*. Thus, the BED-NLR architecture seems to be conserved in this region across grasses, apart from *H. vulgare* (barley) which only showed canonical NLRs in this interval in the studied variety (Morex).

Table 4-11. Number of NLRs in the *Yr* syntenic region across grass genomes including BED-NLRs. BED-I, BED-II and BED-I-II-NLRs are described in Figure 4-11.

Specie	#NLRs	#BED NLRs	#BED NLRs-I	#BED NLRs-II	#BED NLRs-I-II
Rice	6	2	-	-	-
Brachypodium	4	4	*1*	*1*	-
Hordeum vulgare	2	-	-	-	-
***Aegilops tauschii* (D)**	8	4	*1*	*1*	*1*
Hexaploid wheat (D)	6	2	*1*	-	*1*
Wild emmer (A)	8	1	*1*	-	-
Hexaploid wheat (A)	12	5	*3*	*2*	-
Wild emmer (B)	20	10	*6*	*2*	*1*
Hexaploid wheat (B)	13	5	*1*	*1*	*3*

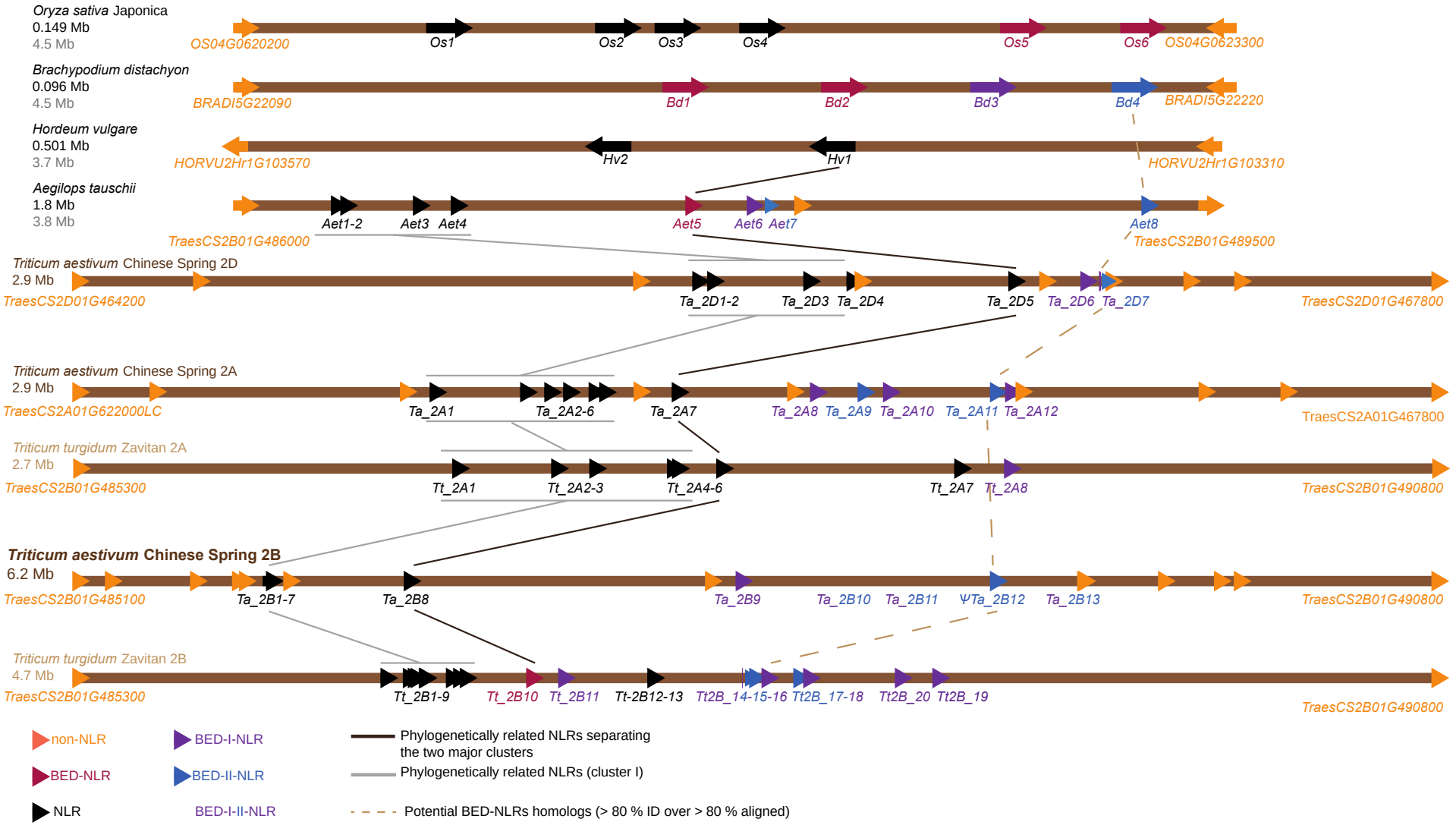

Oryza sativa Japonica
0.149 Mb
4.5 Mb
OS04G0620200
Os1
Os2
Os3
Os4
Os5
Os6
OS04G0623300
Brachypodium distachyon
0.096 Mb
4.5 Mb
BRADI5G22090
Bd1
Bd2
Bd3
Bd4
BRADI5G22220
Hordeum vulgare
0.501 Mb
3.7 Mb
HORVU2Hr1G103570
Hv2
Hv1
HORVU2Hr1G103310
Aegilops tauschii
1.8 Mb
3.8 Mb
Aet1-2
Aet3
Aet4
Aet5
Aet6
Aet7
Aet8
TraesCS2B01G486000
TraesCS2B01G489500
Triticum aestivum Chinese Spring 2D
2.9 Mb
TraesCS2D01G464200
Ta_2D1-2
Ta_2D3
Ta_2D4
Ta_2D5
Ta_2D6
Ta_2D7
TraesCS2D01G467800
Triticum aestivum Chinese Spring 2A
2.9 Mb
TraesCS2A01G622000LC
Ta_2A1
Ta_2A2-6
Ta_2A7
Ta_2A8
Ta_2A9
Ta_2A10
Ta_2A11
Ta_2A12
TraesCS2A01G467800
Triticum turgidum Zavitan 2A
2.7 Mb
TraesCS2B01G485300
Tt_2A1
Tt_2A2-3
Tt_2A4-6
Tt_2A7
Tt_2A8
TraesCS2B01G490800
Triticum aestivum Chinese Spring 2B
6.2 Mb
TraesCS2B01G485100
Ta_2B1-7
Ta_2B8
Ta_2B9
Ta_2B10
Ta_2B11
ΨTa_2B12
Ta_2B13
TraesCS2B01G490800
Triticum turgidum Zavitan 2B
4.7 Mb
TraesCS2B01G485300
Tt_2B1-9
Tt_2B10
Tt_2B11
Tt-2B12-13
Tt2B_14-15-16
Tt2B_17-18
Tt2B_20
Tt2B_19
TraesCS2B01G490800
non-NLR
BED-I-NLR
BED-NLR
BED-II-NLR
NLR
BED-I-II-NLR
Phylogenetically related NLRs separating the two major clusters
Phylogenetically related NLRs (cluster I)
Potential BED-NLRs homologs (> 80 % ID over > 80 % aligned)

Figure 4-10. Expansion of BED-NLRs in the *Triticeae* and presence of conserved BED-BED-NLRs across the *Yr* syntenic region.
Schematic representation of the physical loci containing *Yr7* and *Yr5*/*YrSP* homologs on RefSeq v1.0 and related species. The syntenic region is flanked by conserved non-NLR genes (orange arrows with gene names). Non-NLR genes located in the region are not depicted in Zavitan chromosome. Black arrows represent canonical NLRs and purple/blue/red arrows represent different types of BED-NLRs based on their BED domain and their relationship identified in Figure 4-11 and Figure 4-12. Black lines represent phylogenetically related single NLRs located between the two NLR clusters illustrated in Figure 4-12. Grey lines link NLR-genes from Cluster I, sharing phylogenetic resemblance in the NB-ARC but also sequence similarity across the whole locus. Dashed brown lines show BED-NLRs that are likely to be homologs based on their sequence similarity and location. Details of genes are reported in Appendix 8-15.

4.3.3.2. Identification of two types of BED domains in BED-NLRs belonging to the Yr region in wheat

We hypothesized in Chapter 3 that because the BED domain was nearly identical between Yr7 and Yr5 and identical between Yr5 and YrSP, it does not solely govern resistance specificity. We thus hypothesized that this domain should be conserved across the BED-NLRs located in the *Yr* locus. Based on the analyses above, we asked whether BED-NLRs in this interval had a similar BED domain.

By aligning the BED domains from the identified BED-NLRs in the syntenic region, we discovered two main groups that we subsequently named BED-I and BED-II (Figure 4-10, Figure 4-11). BED-I and BED-II are distinct, with only a few conserved amino-acids that are characteristic of the BED domain in general. These two BED types were found in both hexaploid and tetraploid wheat, *Ae. tauschii* and *B. distachyon*, but not in rice (nor barley which has no BED-NLRs in this interval). In a few instances, we also observed BED-BED-NLRs that were only present in wheat and *Ae. tauschii* (Table 4-11). Interestingly, all BED-BED-NLRs carried BED-I and BED-II in this sequential order. Given that both Yr7 and Yr5 belong to BED-I group (Figure 4-12), we have evidence of this group including functional BED-NLRs.

In all groups (BED-I, BED-II and BED-I-II), the individual BED domain is encoded by a single exon. This leads to two conserved gene structures among the BED-NLRs located in the *Yr* region (Figure 4-11). All BED-I and BED-II share a similar structure harbouring three exons with the BED domain encoded by the second exon (e.g. BED-I *Yr7* and *Yr5*). For BED-I-II, we observed a four exons gene structure with BED-I and BED-II encoded by exon 2 and exon 3, respectively.

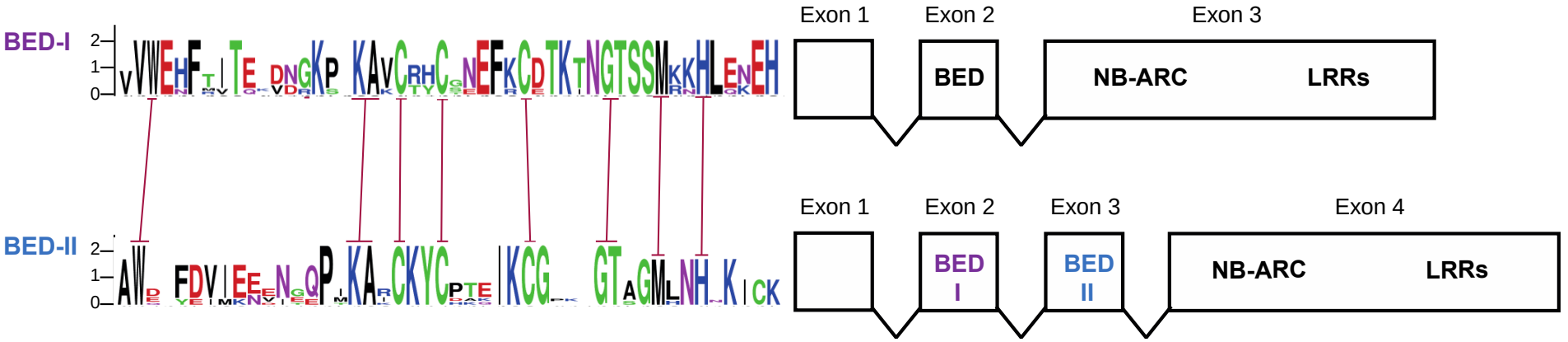


Figure 4-11. Most common gene structure observed for BED-NLRs and BED-BED-NLRs within the *Yr* syntenic interval with associated WebLogo (http://weblogo.berkeley.edu/logo.cgi) diagram showing that the BED-I and BED-II domains are distinct.
Only highly conserved residues defining the BED domain (red bars) were conserved between the two types.

4.3.3.3. Phylogenetic analysis of the NLRs located in the Yr region

Os5 and Os6 are two BED-NLRs from rice that carry BED domains that are too distant to BED-I or BED-II to be categorised as such. However, the gene structure of these loci is identical to what we observed for *Yr7, Yr5* and other BED-NLRs: three exons with the second one encoding the BED domain. We thus investigated whether these rice BED-NLRs were phylogenetically related to the BED-NLRs found in wheat, wild emmer and their progenitor. We constructed a phylogenetic tree based on the NB-ARC domain protein sequence (Figure 4-12). We included the Xa1 protein, that is a BED-NLR providing resistance against bacterial blight in rice[225]. We observed that the canonical NLRs in the syntenic region, which showed a high sequence similarity across the studied grass species (Figure 4-10), are also all phylogenetically linked and belong to one clade (Clade I).

Most BED-NLRs harbouring BED-I and/or BED-II domain belong to a large clade that is distinct from clade I. Between these two clades we observed three small sub-clades, including Clade III that contains the highly conserved NLRs/BED-NLRs across species (back line in Figure 4-10). Interestingly both Os5 and Os6, rice BED-NLRs, are related to this clade. This means they are related to the BED-I and BED-II BED-NLRs. OsXa1 also belong to these subclades, denoting a phylogenetic relationship with wheat BED-NLRs.

From the phylogenetic tree shown in Figure 4-12, we observed that BED-II-containing BED-NLRs share a common ancestor with *B. distachyon* BED-NLRs Bd1 and Bd2. This sub-clade of Clade-II, comprises a mixture of BED-NLRs harbouring BED-I, BED-II or BED-I-BED-II architecture. However, the sub-clade of Clade-II that includes the cloned *Yr* genes is strictly composed of BED-NLRs with the BED-I architecture. Several

hypotheses could explain this. First of all, it is important to note that this tree was computed from the NB-ARC sequences only, thus if the BED domain is under different selection, it could explain why some BED-I are clustering with BED-II NLRs. On the other hand, it could be that the ancestral architecture was BED-I-II and these NLRs there has been differential loss of either BED-I or BED-II in more recent BED-NLRs located in the region. For example, all BED-NLRs in the *Yr* clade would have lost BED-II. Alternatively, the ancestral architecture was BED-I only and there was a second BED domain insertion leading to the BED-I-BED-II organisation, which then diversified into BED-I-NLRs, BED-II-NLRs and some still remained as BED-I-BED-II-NLRs.

From the phylogeny, we observed that BED-NLRs that are not harbouring BED-I or BED-II domains are related to these two different structures based on their NB-ARC (Figure 4-12). The relationship between NLRs from wheat and related species mirrored the phylogeny between these species in general, with rice NLRs being close together in the phylogeny and sharing a distant common ancestor with *Brachypodium*, *Aegilops*, wild-emmer and wheat NLRs. This analysis supports the BED-NLR expansion in wheat and wild-emmer after their divergence from *Brachypodium*. Additionally, Yr5 and Yr7 belong to a strict BED-I-NLR clade. OsXa1 shares a distant common ancestor with this clade. Given that Xa1, Yr7 and Yr5 are functional, we hypothesize that other BED-NLRs that originated from this common ancestral structure could also be functional in grasses.

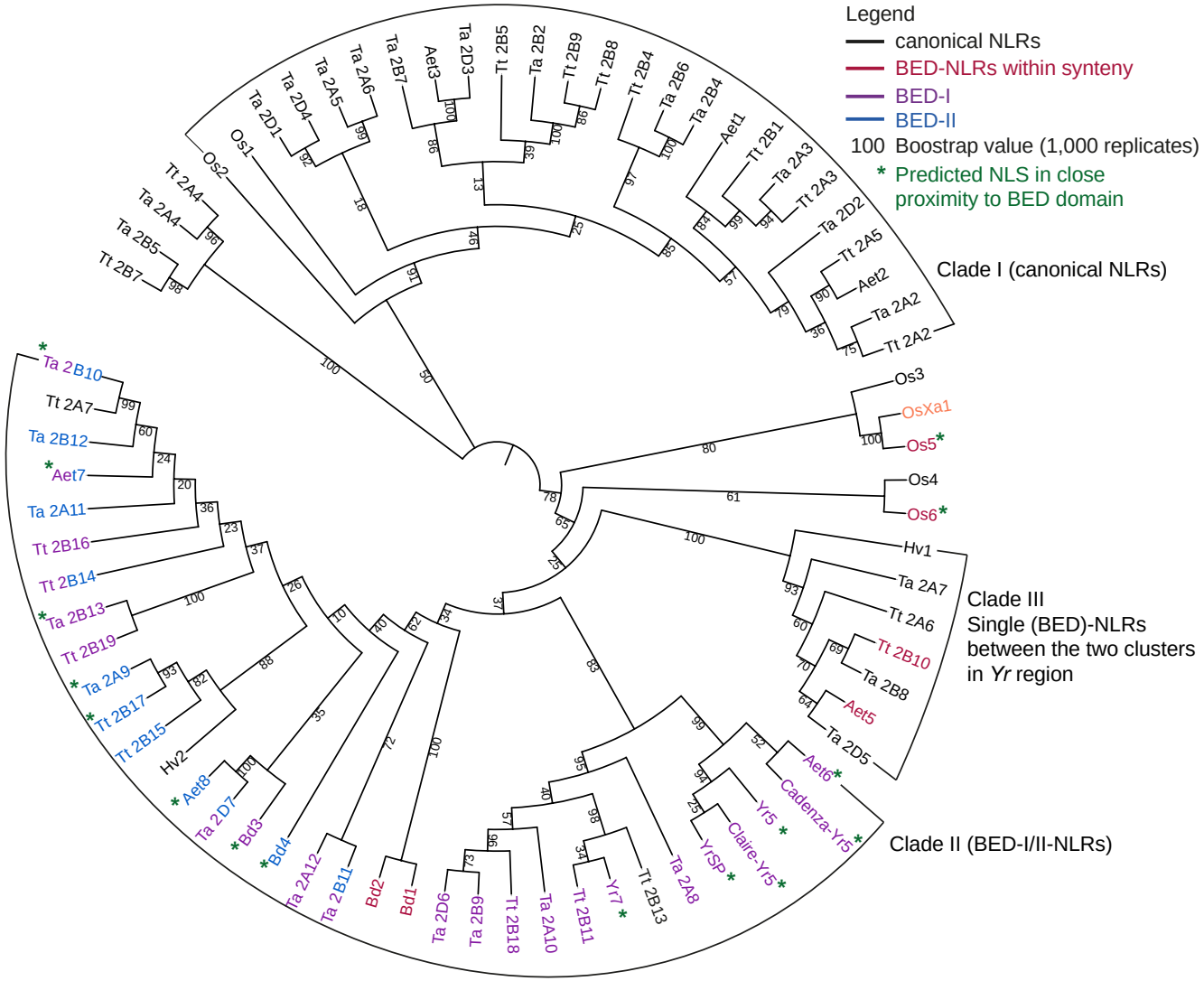


Figure 4-12. The *Yr* loci are phylogenetically related to nearby NLRs on RefSeq v1.0 and their orthologs. Phylogenetic tree based on translated NB-ARC domains from NLR-Annotator. Node labels represent bootstrap values for 1,000 replicates. The tree was rooted at mid-point and visualized with Dendroscope v3.5.9. The colour pattern matches that of Figure 4-10 to highlight BED-NLRs with different BED domains

4.3.3.4. Identification of a Nuclear Localisation Signal in Yr7 and Yr5

Nuclear Localisation Signals (NLS) were found in previously cloned rice BED-NLRs Xa1 and the Xo1 candidate[226]. Given that Xa1 was distant but still phylogenetically related to Yr7 and Yr5, we asked the question whether the presence of an NLS would be an additional feature of BED-NLRs. Given that the NLS were located in close proximity downstream the BED domain in Xa1 and Xo1 candidate, we extracted the extended region surrounding the BED domain (ten amino-acid upstream and 60 downstream) in BED-NLRs located in the *Yr* region and used NLSdb[267] (https://rostlab.org/services/nlsdb/) to predict NLS.

We confirmed the presence of an NLS in the vicinity of the BED domain (~ 30 amino-acids downstream) in a subset of the BED-NLRs carrying a single BED domain located in the syntenic region in wheat (Figure 4-13). BED-NLRs possessing two BED domains back to back also possess NLS, located after the first domain (Figure 4-13). As previously stated, the first BED domain is similar to BED-I domain from single BED-NLRs and second BED domain is similar to BED-II.

Although the NLS are different and not located at the same position in Xa1 and Yr7/Yr5, this feature is present in the only BED-NLR immune receptors that have been shown to be functional in rice and wheat. We will test the functionality of Yr7 NLS in Chapter 5 to determine whether it has the ability to transfer the Yr7 protein into the nucleus. More work will be required to confirm that the NLS feature is required for the expression of resistance *in planta*.

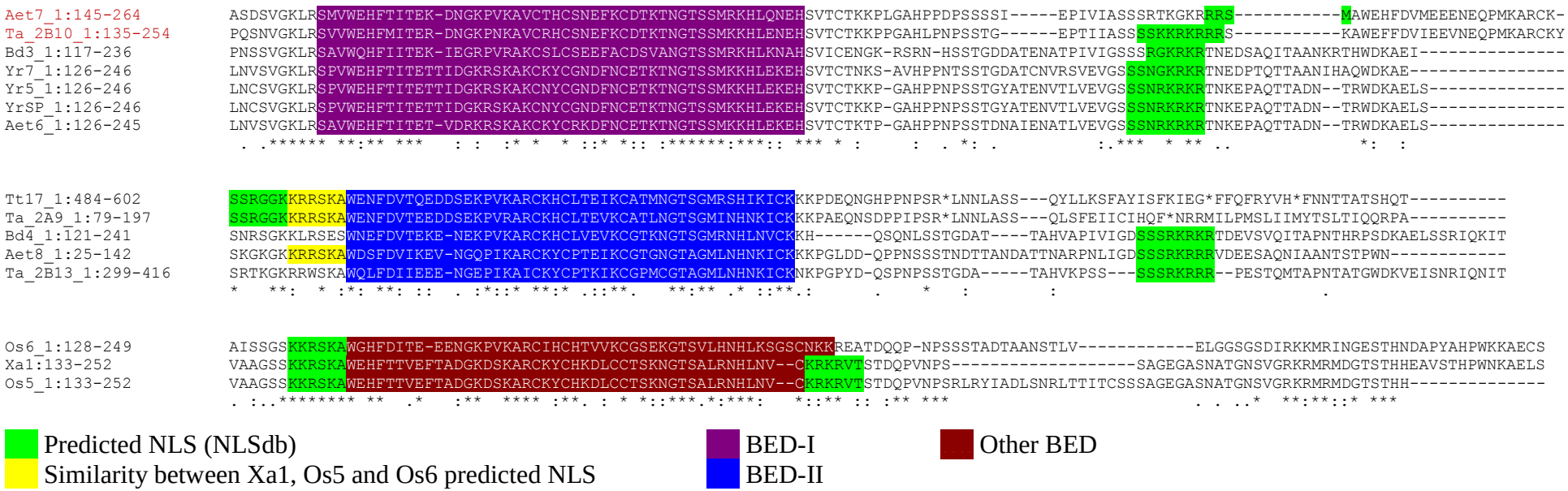


Figure 4-13. Alignment of the region surrounding the predicted BED domain (10/+100 amino-acids) in BED-NLRs containing predicted NLS (NLSdb). BED NLRs in red are BED-I-II NLRs.

In this section we provide evidence that the *Yr* locus has expanded in wheat, wild-emmer and *Ae. tauschii*, although BED-NLRs were also present in this region in both *B. distachyon* and *O. sativa* (Figure 4-10, Table 4-11). These BED-NLRs were phylogenetically related to Yr7 and Yr5 (Figure 4-12). We identified two conserved types of BED domains in wheat, wild-emmer and *Ae. tauschii* (BED-I and BED-II, Figure 4-11). Both Yr7 and Yr5 contain BED-I BED domains so there is evidence for this sequence to be functional. We found NLS in a subset of the BED-NLRs present in the *Yr* region, including Yr7 and Yr5. Two NLS are present in Xa1. We thus hypothesized that this feature was important for protein function, which will be explored in Chapter 5.

4.3.4. Neighbour-network analysis of BED domains from BED-NLRs and from other BED-containing proteins in wheat

We established that BED-NLRs are a conserved gene structure in the *Yr* region in wheat and related grasses. In Chapter 3, we hypothesized that BED domains could act as an integrated domain important for effector recognition. The integrated decoy model, presented in section 4.1.1, proposes that the NLR with an integrated domain (NLR-ID) recognises the pathogen either by direct binding of an effector or due to protein modification activity of the effector on the integrated domain[73,211]. In both cases, it implies that the integrated domain has similar sequence and/or structure to the initial target of the effector. Hence, we hypothesized that identifying such conserved features between BED domains from NLRs and from other proteins could inform us on the pathogen target.

To investigate this, we identified all BED-containing proteins in RefSeqv1.0, extracted their BED domain and conducted a Neighbour-net analysis to determine whether BED

domains from BED NLRs shared common features with BED domains from other non-NLR BED-containing proteins (Figure 4-14). We also retrieved the rice Xa1 and ZBED proteins, the latter being hypothesized to mediate rice resistance to *M. oryzae*[87]. Overall, BED domains are diverse, although there is evidence of a split between BED domains from BED-NLRs and non-NLR proteins. However, there was 7 of 83 BED domains from non-NLRs clustering with the BED domains from BED-NLRs.

Interestingly, five of these seven proteins also carry a hAT family C-terminal dimerisation domain (PF05699.9. Table 4-12, Appendix 8-16). This domain could either belong to a functional or domesticated hAT transposase, such as the daysleeper proteins[236] we mentioned in Chapter 3. We conducted a BLASTp analysis against the NCBI database and most of these proteins had a ricesleeper-like protein as a best hit (Appendix 8-16). Further investigation will be required to decipher whether these proteins are potential target of the effectors recognized by BED-NLRs during infection.

Table 4-12. Record of the additional domains in proteins whose BED domain clusters with BED domains from BED-NLRs (Figure 4-14)

Protein ID (RefSeqv1.0)	Additional domain
TraesCS3B01G269600.1	Dimer_Tnp_hAT
TraesCS7A01G447400.1	Dimer_Tnp_hAT
TraesCS3B01G317800.1	Dimer_Tnp_hAT
TraesCS5B01G501500.1	Oxygenase-NA
TraesCS5D01G501900.1	Oxygenase-NA
TraesCS5B01G377100.1	Dimer_Tnp_hAT/Acetyltransf_1
TraesCS1B01G158800.1	Dimer_Tnp_hAT

Given that the base of the split is broad on Figure 4-14, integrated BED domains are diverse and may also have derived from multiple integration events, although Yr7 and Yr5/YrSP both arose from a common integration event that occurred before the Brachypodium/wheat divergence (Figure 4-14). The split between BED domains from BED-NLRs and BED domains from BED-containing proteins is consistent with the

hypothesis that integrated domains might have evolved to strengthen the interaction with pathogen effectors after integration[268]. Hence the evolution constraint applied to integrated domains would be different than of the initial target. However, we still cannot exclude the potential role of the BED domains in signaling at this stage.

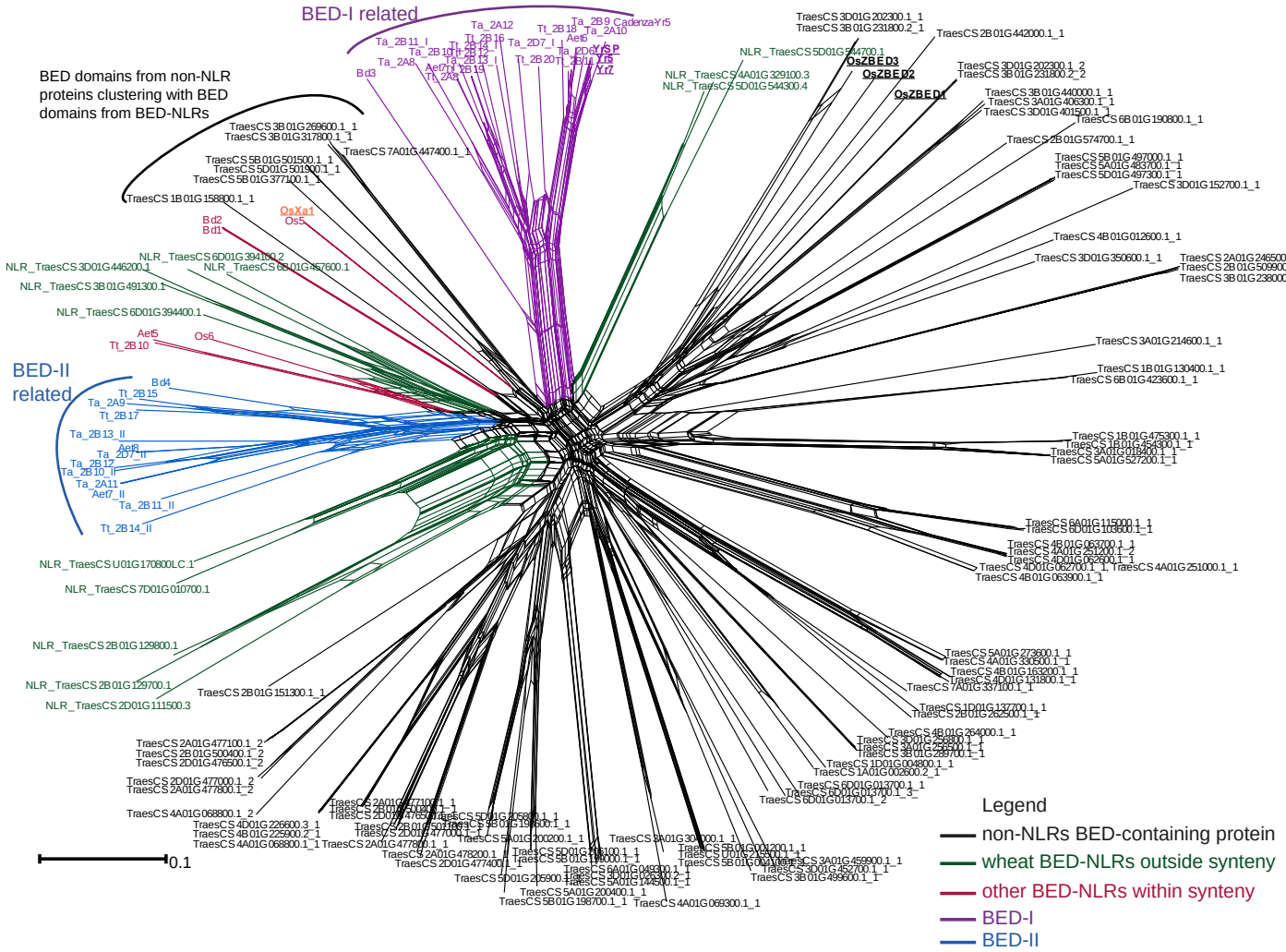


Figure 4-14. Neighbour-net analysis based on uncorrected P distances obtained from alignment of 153 BED domains including the 108 BED-containing proteins (including 25 NLRs) from RefSeq v1.0, BED domains from NLRs located in the syntenic region as defined in Figure 4-10, and BED domains from Xa1 and ZBED from rice.

BED-I and II clades are highlighted in purple and blue, respectively. BED domains from the syntenic regions not related to either of these types are in red. BED domains derived from non-NLR proteins are in black and BED domains from BED-NLRs outside the syntenic region are in grey.

4.3.5. Re-analysis of a RNA-seq time-course during *Pst* infection

We investigated previously published RNA-Seq data from a *Pst* infection time-course of the cultivars Vuka and AvocetS-Yr5 (*Pst* isolate 87/66, virulent to Vuka and avirulent to AvocetS-Yr5). Our aim was to identify the expression patterns of BED-containing proteins in this context. We hypothesized that providing some BED-containing protein might be the target of *Pst* effectors, such targets should be expressed in the susceptible cultivar under infection to allow for interaction with the effector.

There were only two BED-containing loci identified as differentially expressed at any given time point after 0 dpi between Vuka and AvocetS-Yr5 (TraesCS2A01G477800 and TraesCS2D01G477000) (Table 4-13). The two proteins carried domain of unknown function (DUF) and dimerization region is found at the C terminus of the transposases of elements belonging to the Activator superfamily (hAT element superfamily, PF05699). However, the basal expression values were very low ($\leq$ 1 TPM, Table 4-13) so it is unsure whether this difference in the expression is biologically relevant.

Table 4-13. Summary of the only two BED-containing proteins found differentially expressed at any time point after 0 dpi between AvocetS-Yr5 and Vuka (adjusted *p*-value < 0.05)

Gene model	Vuka 1dpi (TPM)	AvocetS-Yr5 1dpi (TPM)	Log2 fold change	Adjusted *p*-value
TraesCS2A01G477800	0.68	0.33	4.07	0.006
TraesCS2D01G477000	1.16	0.87	5.2	5.19E-7

The expression values of BED-NLRs and BED-containing proteins (based on RefSeqv1.0 gene models) is shown as a heatmap in Figure 4-15. Most of the BED-proteins are not expressed, including many BED-NLRs. Those that are expressed have an overall low expression value and no obvious pattern could be observed relating to the presence of *Pst*. Additionally, the seven BED-proteins clustering with BED-NLRs on

Figure 4-14 (in green on Figure 4-15) were not expressed at all during *Pst* infection. This thus does not support our hypothesis proposing that certain BED-containing proteins would be highly expressed in Vuka as compared to AvocetS-Yr5 during *Pst* infection. However, regarding the expressed BED-containing proteins in both varieties, this is consistent with the prediction that effectors alter their targets' activity at the protein level in the integrated-decoy model[73], rather than at the transcriptional level. We thus cannot disprove that BED-containing proteins are involved in BED-NLR mediated resistance.

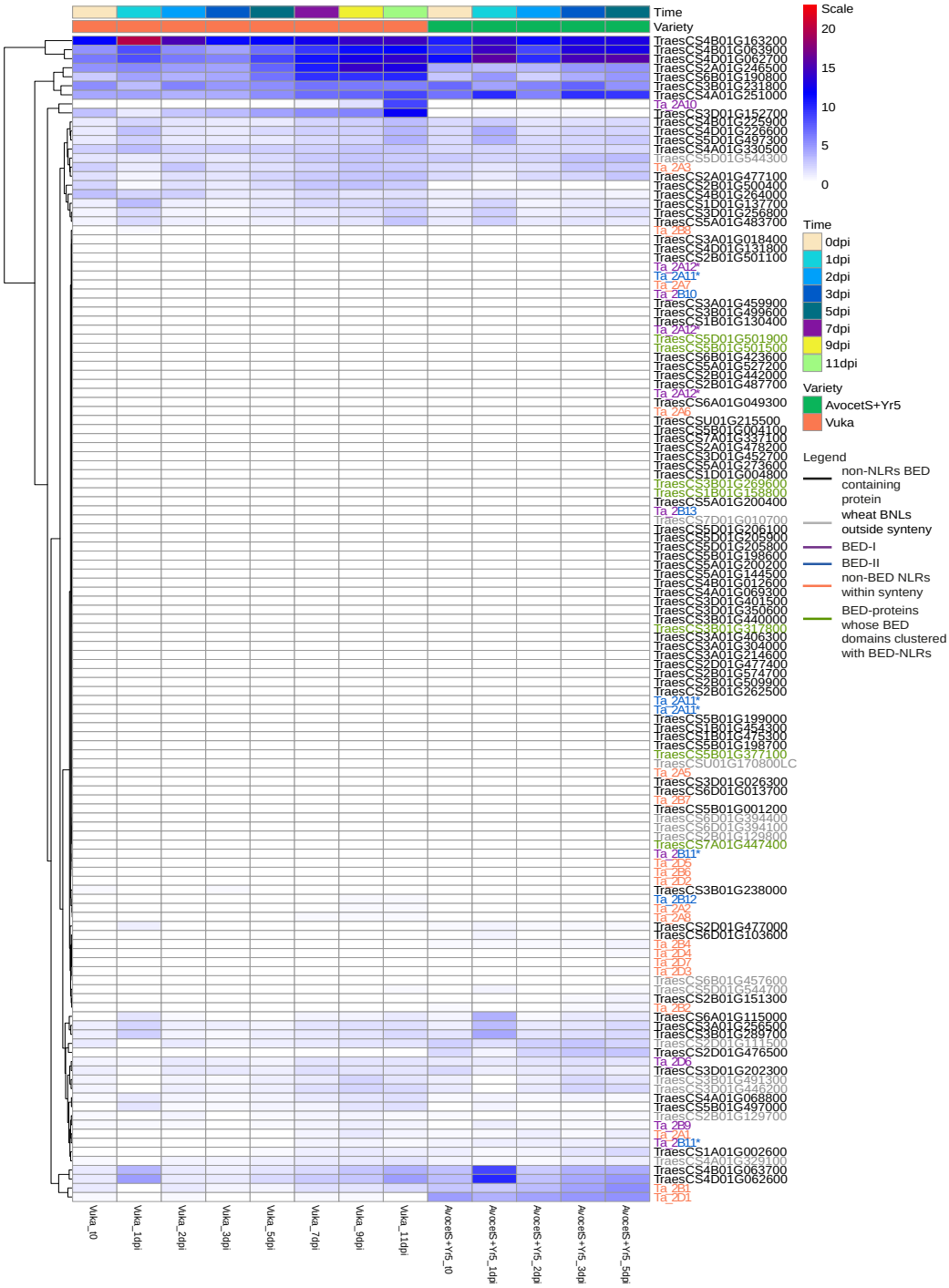


Figure 4-15. Heatmap representing the normalised read counts (Transcript Per Million, TPM) from the reanalysis of published RNA-Seq data[197] for all the BED-containing proteins, BED-NLRs and canonical NLRs located in the syntenic region annotated on RefSeq v1.0 during yellow rust-infected time-course in susceptible and resistant cultivars.

The RefSeq gene model identifier was replaced with the one used in the synteny figures for BED-NLRs to allow for comparison. Lack of expression is shown in white and expression levels increase from blue to red. Asterisks show cases where several gene models were overlapping with NLR loci identified with NLR Annotator. The colour pattern matches that of Figure 4-14 to highlight BED-NLRs with different BED domains. Orange labels show the expression of the canonical NLRs located within the syntenic interval (black on Figure 4-10). The seven non-NLR BED genes whose BED domain clustered with the ones from BED-NLR proteins in Figure 4-14 and are indicated in green.

4.3.6. Neighbour-network analysis of BED domains from BED-NLRs and from other BED-containing proteins in plants

We observed a split between BED domains from BED-NLRs and BED domains from other BED-containing proteins in wheat, although the base of the split was very broad and thus suggested high amount of variation among BED domains in general (Figure 4-14). We hypothesized that finding a similar pattern in other plants might strengthen our observation from wheat. We thus queried for BED-NLRs in plant proteomes that were available on Phytzome (https://phytozome.jgi.doe.gov/pz/portal.html) and EnsemblPlants (https://plants.ensembl.org/index.html) and kept only the ones showing an acceptable BUSCO score (Appendix 8-17).

4.3.6.1. Identification of BED-NLRs in deposited plant proteomes

We investigated a total of 90 proteomes from 84 species. We used the BUSCO program[262] to estimate whether the protein sets were complete. BUSCO stands for Benchmarking Universal Single-Copy Orthologs and uses evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB v10[263] (http://www.orthodb.org/) to determine completeness of a proteome. Given that we had proteomes ranging from algae to Angiosperms, we used two different orthologs sets for the analysis: Viridiplantae (430 orthologs, includes algae) and Embryophytes to refine the analysis on plants (1,440 orthologs). We selected proteomes showing at least 90 % of completeness in both cases, or only in Viridiplantae analysis for species that do not belong to Embryophytes. This resulted in a working set composed of 68 proteomes (69 including RefSeqv1.0).

We screened the proteomes for BED domain using the hmmer program (section 4.2.5). In total, 65 out of 68 proteomes contained BED domains in their proteins (66/69 including RefSeqv1.0). This domain is thus frequent in plants. However, only 18 of the 66 proteomes contained a NB-ARC domain in addition to the BED domain within the same protein. This includes grasses, as we expected from our previous analysis, but also members of the Fabidae and Malpighiales (Figure 4-16). Note that we lack *Aegilops tauschii*, *Triticum dicoccoides* and *Leersia perrieri* species from the tree presented on Figure 4-16 as these species were obtained from sources different from Phytozome (Appendix 8-17). From the tree we can observe that BED-NLRs are an Angiosperm innovation, although BED containing proteins are present in Chlorophytes (Appendix 8-17). Additionally, the analysis would suggest that BED-NLRs are derived from several independent integration events; alternatively most eudicots have lost them (Figure 4-16).

We observed on Table 4-14 that the proportion of BED-NLRs among the total identified BED-proteins varies between plant species, even within phylogenetically closely related species. *Triticum turgidum* ssp. *dicoccoides* and *Triticum aestivum* displayed the highest BED-NLR/total BED-proteins ratios. Other species from the large grass clade also displayed ~ 20 % BED-NLRs (e.g *Leersia perrieri*, *Setaria italica*, *Aegilops tauschii*, *Brachypodium distachyon*). This was similar to what we observed for the Malpighiales *Populus trichocarpa* and *Salix purpurea*. The Fabidae showed low number of BED-NLRs in general, similar to what we observed in rice. Additionally, despite showing a very high number of BED-containing proteins, only a few BED-NLRs (0.8 to 3.8 %) were recorded for *Panicum virgatum*, *Panicum halli* and *Zea mays* (Table 4-14). There was thus no obvious pattern within phylogenetically close species regarding BED-NLR content.

Table 4-14. Summary of the species containing BED-NLRs in their proteomes and proportion of the BED-NLRs in respect to the total BED-proteins.
Coloured groups correspond to phylogenetically close species and we performed a split-network analysis similar to what we showed on Figure 4-14 for each of these groups.

Species	#BED-NLRs	#Total BED-containing proteins	%BED-NLRs
Brachypodium stacei	1	9	11.1
Hordeum vulgare	4	23	17.4
Aegilops tauschii (High confidence proteins only)	8	44	18.2
Brachypodium distachyon	4	18	22.2
Triticum aestivum (High confidence proteins only)	22	37	59.5
Triticum dicoccoides	12	20	60.0
Zea_mays	1	123	0.8
Panicum hallii	2	74	2.7
Panicum virgatum	9	234	3.8
Setaria italica	2	10	20.0
Oryza sativa indica	2	29	6.9
Oryza sativa japonica	4	49	8.2
Leersia perrieri	3	13	23.1
Glycine max	3	136	2.2
Eucalyptus grandis	1	31	3.2
Trifolium pratense	3	49	6.1
Phaseolus vulgaris	2	8	25.0
Medicago truncatula	2	3	66.7
Populus trichocarpa	20	90	22.2
Salix purpurea	19	62	30.6

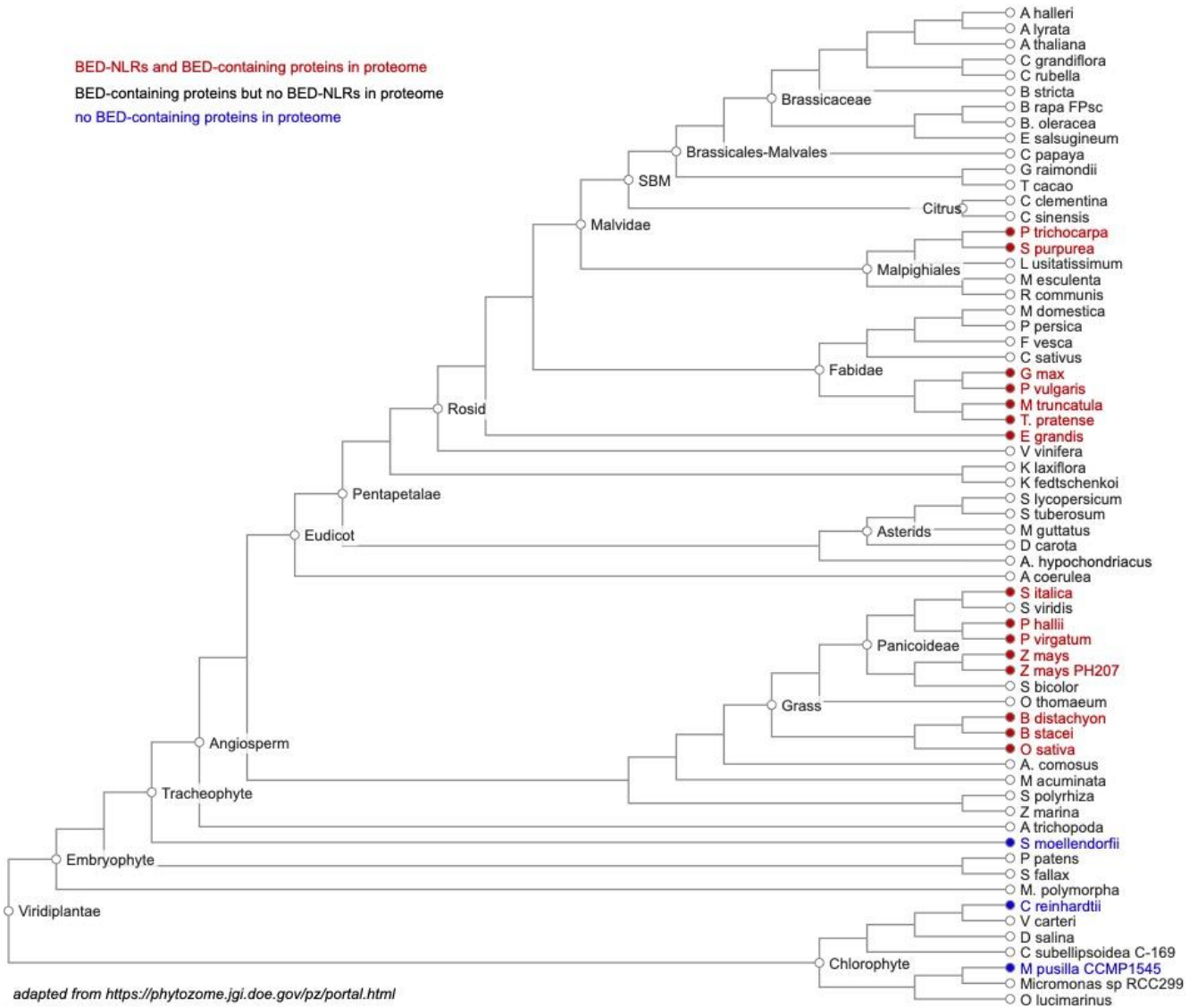


Figure 4-16. Phylogenetic tree represented the plant species whose genomes are on Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html).

No modification was made to the structure of the tree, we highlighted in red the species in which we found BED-NLRs and BED-containing proteins, in black species with BED-containing proteins but not BED-NLRs and in blue species in which we did not find any BED domain.

4.3.6.2. Split network analysis in plant proteomes containing BED-NLRs

We carried out the same analysis as in Figure 4-14 in phylogenetically close plant species carrying BED-NLR proteins to determine whether BED domains from BED-NLRs shared similarities with BED domains from other proteins. Identifying similarities between BED domains could inform on the identity and role of the potential effector target, within the integrated decoy model.

Pooideae (orange group in Table 4-14, Figure 4-17)

We defined four major clades (I to IV) in the neighbour-net analysis performed on bread wheat (*Triticum aestivum*), emmer wheat (*Triticum dicoccoides*), goat grass (*Aegilops tauschii*) and barley (*Hordeum vulgare*) and Brachypodium (*Brachypodium distachyon* and *Brachypodium stacei*). We adopted an inclusive approach to define the clades, even if there was early diverge at the base (see example of Clade-II in the following paragraph). Clade-I was the clade comprising the most BED-domains from BED-NLRs (32) and only six from other BED-proteins. Moreover, all the clustering BED-proteins were short and only carried a BED domain (no additional domain, Appendix 8-18), suggesting that these could be incomplete annotation.

Clade-II was composed of comparable numbers of BED-NLRs and other non-NLR BED-proteins (eleven and nine, respectively, Figure 4-17). However, both groups diverged early in the clade, depicting sequence variation between BED-NLRs and other BED-proteins (Figure 4-17). Conserved domain organization of the full BED-proteins from which the BED domains are derived from was similar to what was observed in Clade-I (no additional domain found, Appendix 8-18). A similar profile was observed for Clade-III (four BED-NLRs and 15 BED-containing proteins, Figure 4-17).

Interestingly, 4/11 BED-proteins clustering with BED-NLRs from Clade-IV carried additional Dimer_Tnp_hAT and DUF4413 domains, as identified on Table 4-12. The Yr7-Clade was highly divergent (note that only high confidence gene models from RefSeqv1.0 were used here, hence the Yr7-like BED domain clade is smaller than on Figure 4-14). Indeed, its members do not share many links apart from at the very base (Figure 4-17). Among the additional domains that BED-proteins in this clade carry, we observed 2/18 Dimer_Tnp_hAT and DUF4413, 2/18 DnaJ (member of the hsp40 family of molecular chaperones) and NAM (no apical meristem domain) and 13/18 single BED domains.

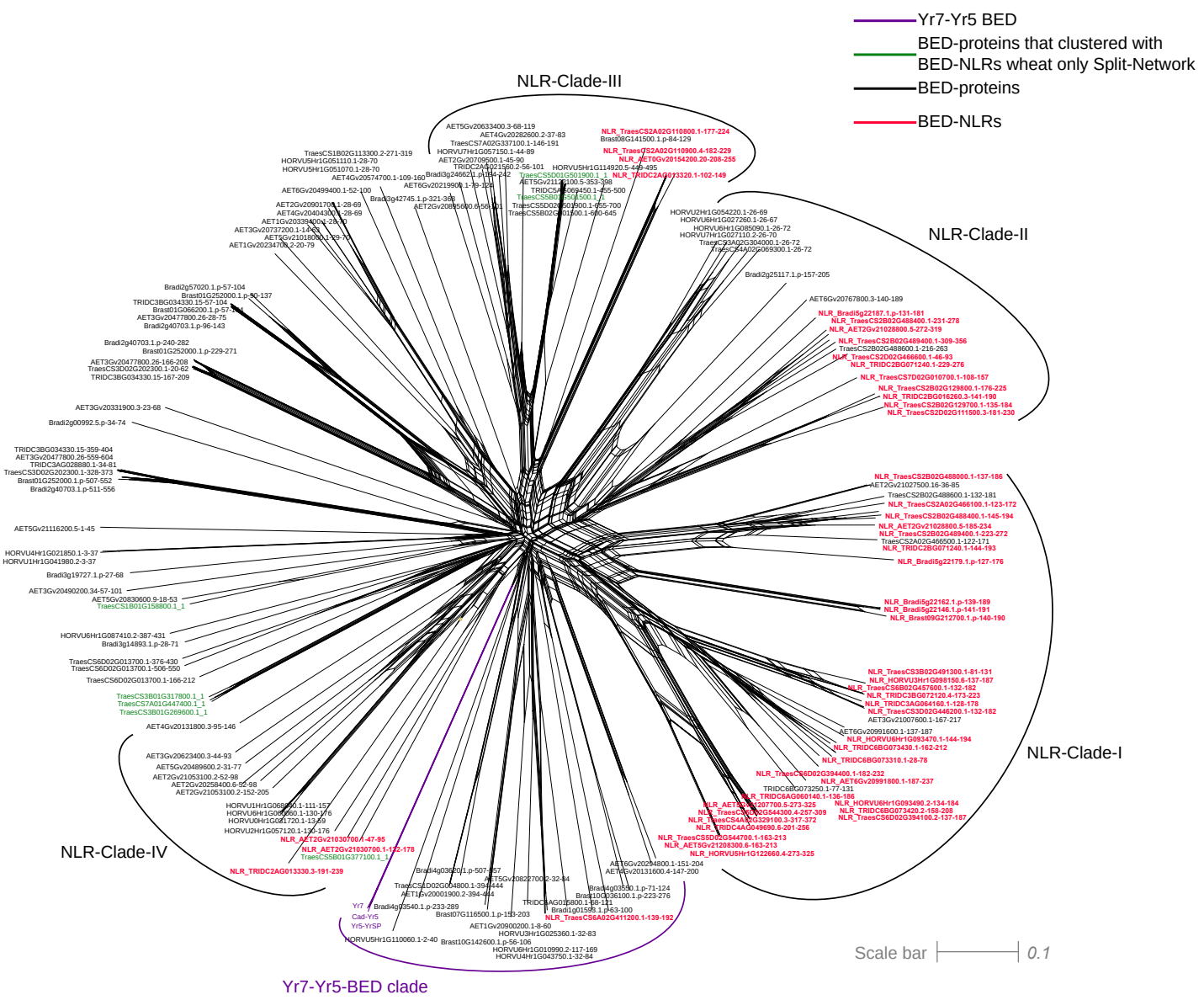


Figure 4-17. Neighbour-net analysis based on uncorrected P distances obtained from alignment of 151 BED domains including 51 BED- NLRs from the orange group defined on Table 4-14 (Pooideae).

BED-NLR are shown in pink and other BED-containing proteins in black. Yr7 and Yr5 BED domains are shown in purple (BED-I colour on Figure 4-14). We added in green the BED-proteins that we found clustering with BED-NLRs in wheat. We used an updated annotation (RefSeqv1.1 High Confidence genes only) *Triticum aestivum* in this figure and not all the previously annotated BED-containing proteins were present in this new version.

Overall these observations confirm what we reported for wheat in Figure 4-17. Thus, this can be generalised to the species most closely related to wheat. Indeed, most of the clades harboured members of the six tested proteomes. We found that BED-proteins clustering with BED-NLRs were mostly single-BED domain proteins, with few occurrences of Dimer_Tnp_hAT and DUF4413.

Ehrhartoideae (brown group in Table 4-14, Figure 4-18)

We observed in Figure 4-14 that Xa1 BED domain was distantly related to both BED-I and BED-II domains. We thus carried out a similar analysis as above but focusing on two *Oryza sativa* sub-species (indica and japonica) and a distantly related specie *Leersia perrieri* to assess this in greater detail. Compared to the Pooideae analysis (orange group), there are less BED-proteins and fewer BED-NLRs in this Erhartoideae group (Table 4-14).

Clade-I only comprised five BED-NLRs and two BED-proteins, both single BED proteins (Appendix 8-18). An opposite composition was observed in Clade-II, where 10 BED-proteins clustered with only two BED-NLRs. More than half of these BED-proteins were single BED proteins (6/10), three were Dimer_Tnp_hAT proteins with additional DUF (domain of unknown function) and the last one only carried a DUF domain (Appendix 8-18). A comparable composition was observed in the Xa1 clade (twelve BED-proteins and three BED-NLRs, Appendix 8-18), including a BED-protein carrying a Fbox (protein-protein interaction) domains. The Xa1 clade was highly divergent, similarly to what we observed for the Yr7 clade in wheat (Figure 4-18).

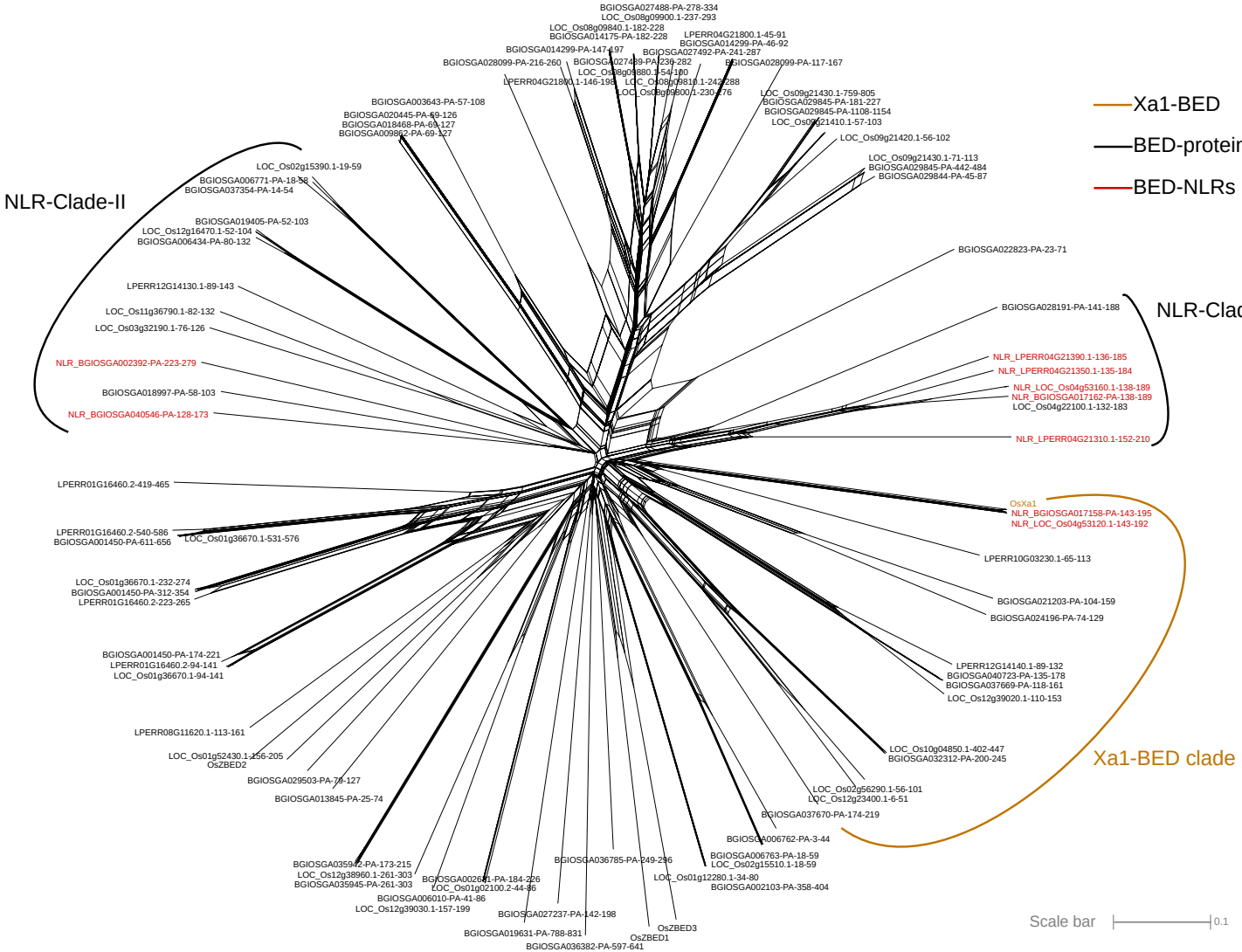


Figure 4-18. Neighbour-net analysis based on uncorrected P distances obtained from alignment of 91 BED domains including 9 BED- NLRs from the brown group defined on Table 4-14 (Ehrhartoideae).

BED-NLR are shown in pink and other BED-containing proteins in black. Xa1 BED domains and its clade are shown in dark orange.

Panicoideae (yellow group in Table 4-14, Figure 4-19)

We analysed together *Panicum virgatum*, *Panicum halleri*, *Zea mays* and *Setaria italica* (Figure 4-19). *Panicum virgatum*, *Panicum halleri* and *Zea mays* were among the species having the lowest ratios of BED-NLRs/BED-proteins in the tested proteomes, despite having a higher number of BED-proteins when compared to the other species (Table 4-14).

Interestingly, BED-NLRs clustered in three separate clades despite their very low absolute numbers (Figure 4-19). Clade-I comprises 62 BED-proteins clustering with only two BED-NLRs. Most of these proteins were single BED proteins (47/62). We also identified 11/62 Dimer_Tnp_hAT with a DUF. Two additional proteins only encoded DUFs, one carried a Myb_DNA-binding domain (Appendix 8-18) and the last one had five additional domains (DUF4413, Dimer_Tnp_hAT, NPR1_like_C, Ank, DUF3420). Clade-II has nine BED-NLRs and 13 BED-proteins and, seven of which were single BED proteins and two double BED proteins (Appendix 8-18). The remaining proteins carried DUFs, including one with a Dimer_Tnp_hAT domain. The last protein encoded two BED domains and a NAM domain. Clade-III contained three BED-NLRs and 23 BED-proteins in total, of which eight were single BED domain proteins, ten were Dimer_Tnp_hAT, four were Fbox containing proteins and the last one carried a transmembrane domain (Appendix 8-18).

The expansion of the number of BED-proteins in *Panicum virgatum*, *Panicum halleri* and *Zea mays* did not seem to have an influence on the nature of BED-proteins whose BED domains share similarities with BED-NLRs. Indeed we observed the same associated domains (Dimer_Tnp_hAT, DUFs, signle-BED) than in the Pooideae (orange) and Ehrhart oideae (brown). Additionally, Clade-II comprises most of the

Panicum virgatum BED-NLRs (6/9), which is similar to what we observed in Clade-I in Figure 4-17 in wheat and related species.

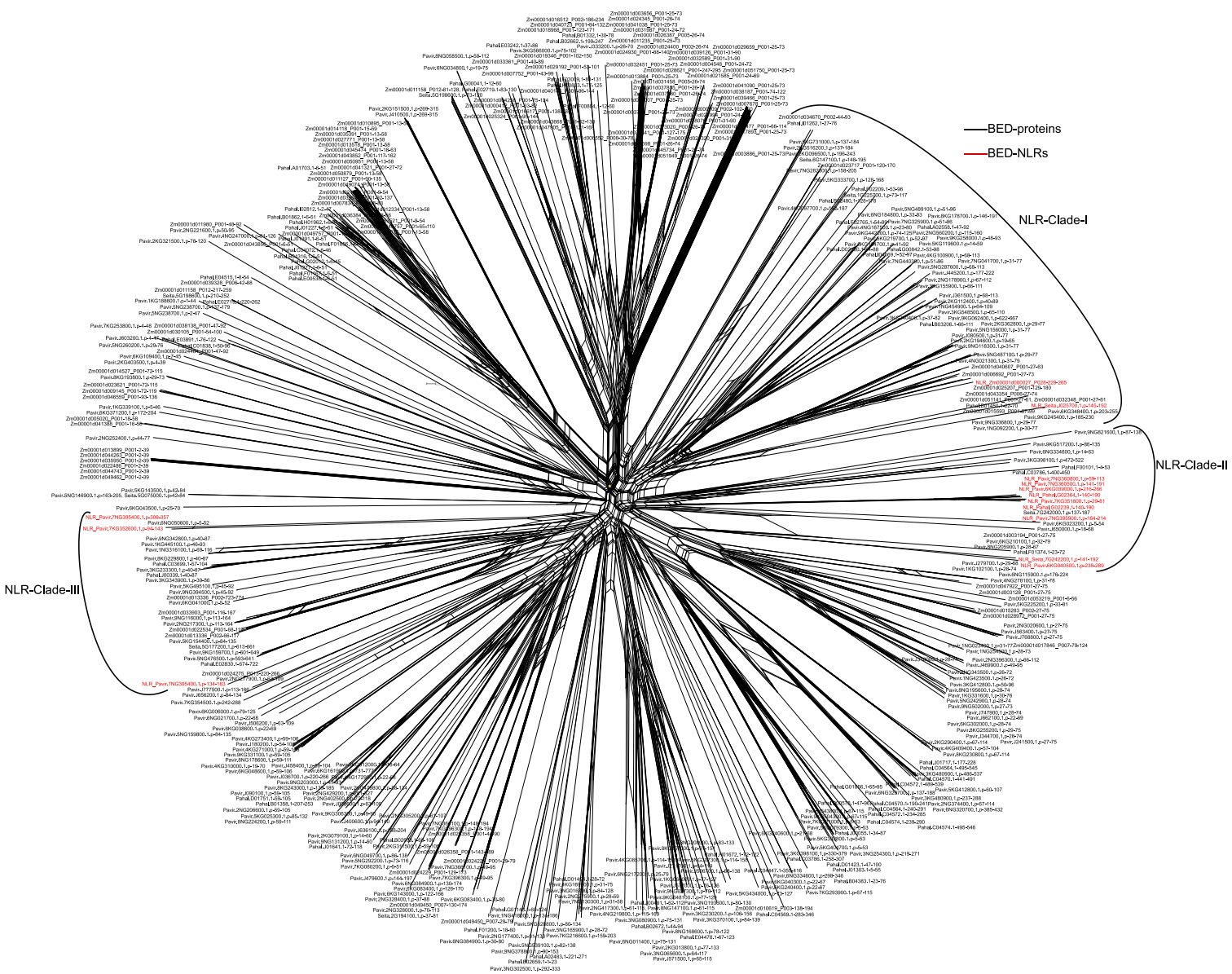


Figure 4-19. Neighbour-net analysis based on uncorrected P distances obtained from alignment of 441 BED domains including 14 BED- NLRs from the yellow group defined on Table 4-14 (Panicoideae).

BED-NLR are shown in red and other BED-containing proteins in black.

We observed a certain conservation in the domain architecture of BED-proteins whose BED domain share similarities with BED-NLRs in monocots. We thus pursued the same analysis in eudicots that harboured BED-NLRs to determine whether a similar trend could be observed (Figure 4-16).

Fabideae and *Eucalyptus grandis* (blue group in Table 4-14, Figure 4-20)

We identified BED-NLRs in *Glycine max*, *Phaseolus vulgaris*, *Medicago truncatula* and *Trifolium pratense* (Table 4-14). We added *Eucalyptus grandis* to this group given its position on the tree in Figure 4-16 and that it displays BED-NLR/BED-protein ratios similar to *Trifolium pratense.* Overall, there was very few BED-NLRs identified in this group (10 BED-NLRs for 227 BED-proteins in total). *Glycine max* has a BED-NLR/BED-protein ratio similar to what we observed in *Panicum virgatum*, *Panicum halleri* and *Zea mays*, that is an expansion of the number of BED-protein when compared to the other proteomes, but very few BED-NLRs among them. Both *Phaseolus vulgaris* and *Medicago truncatula* showed low numbers of total BED-proteins (8 and 3, respectively, Table 4-14).

We defined a total of four clades containing BED-NLRs and other BED-proteins (Figure 4-20). Clade-I was the largest and most clearly defined clade, with the BED-proteins branching from the BED-NLRs. We reported two BED-NLRs and 45 BED-proteins in this clade, all but two derived from *Glycine max*, and the majority (37) encoded Dimer_Tnp_hAT coupled with a DUF domain (Appendix 8-18). Four proteins encoded DUFs only, one was a single-BED protein and the last one harboured a SWIB domain (first identified in a protein involved in chromatin remodelling).

Clade-II comprised eight BED-proteins clustering with one BED-NLRs from *Trifolium pratense* (Figure 4-20). All BED-proteins but one, were derived from this plant species. We reported three Dimer_Tnp_hAT coupled with a DUF domain, three single BED domain proteins and, one protein encoding a single DUF and the last one harbouring a DBD_Tnp_Hermes (similar to hAT but from another transposon family) (Appendix 8-18).

Clade-III also had eight BED-proteins clustering with four BED-NLRs. The BED-NLRs derived from *Eucalyptus grandis* (1), *Trifolium pratense* (1) and *Phaseolus vulgaris* (2) and the BED-proteins originated from *Eucalyptus grandis* (4), *Trifolium pratense* (1) and *Glycine max* (3). The four BED-proteins derived from Eucalyptus were all double-BED domain proteins and the other were Dimer_Tnp_hAT coupled with a DUF proteins (Appendix 8-18).

Clade-IV showed an equal number of BED-NLRs and BED-proteins (3, Figure 4-20). BED-NLRs derived from *Glycine max* and *Medicago truncatula* and the BED-proteins derived from *Glycine max* (single-DUF proteins) and *Trifolium pratense* (single-BED proteins).

It is surprising to observe that no BED-proteins from *Medicago truncatula* and *Phaseolus vulgaris* clustered with BED-NLRs from the same species. This could mean that either these species lost these corresponding BED-proteins, or they are diverged from the BED domains from BED-NLRs to an extent that BED-proteins from related species are more similar. We recorded the combination of Dimer_Tnp_hAT domain coupled with DUF proteins, as well as single-BED proteins, carrying BED domains that are similar to those of BED-NLRs. This is consistent with what to observed for the previous groups.

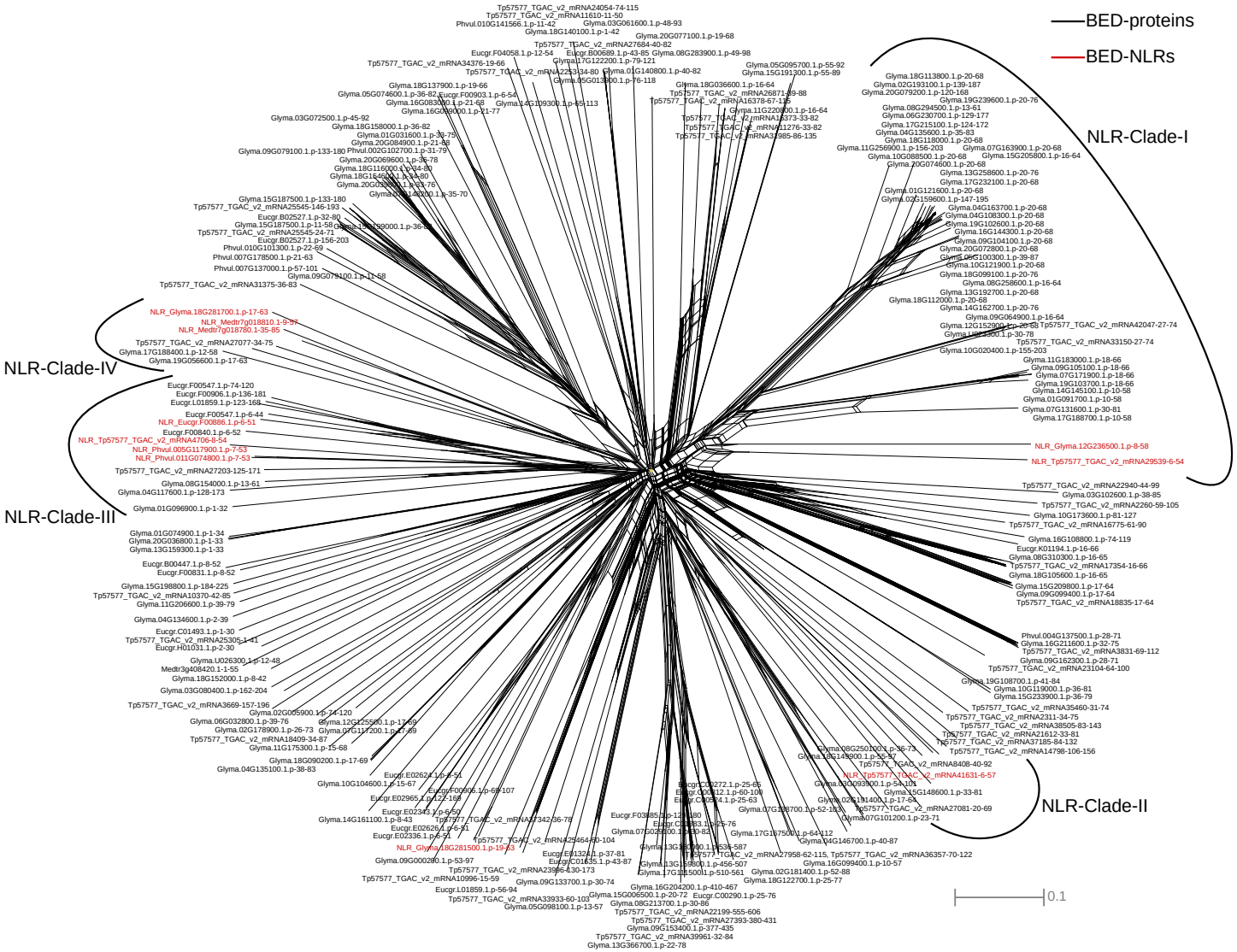


Figure 4-20. Neighbour-net analysis based on uncorrected P distances obtained from alignment of 227 BED domains including 10 BED- NLRs from the blue group defined on Table 4-14 (Fabideae + *Eucalyptus grandis*).

BED-NLR are shown in red and other BED-containing proteins in black.

Malpighiales (green group in Table 4-14, Figure 4-21)

Populus trichocarpa and *Salix purpurea* harboured a BED-NLRs/BED-proteins ratio similar to what we observed in *Aegilops tauschii*, *Brachypodium distachyon*, *Setaria italica*, *Leersia perrieri* and *Phaseolus vulgaris* (~ 20-30 %, Table 4-14).

We defined one clear clade (Clade-I) and three other sub-clades (Clade-II to IV) that we will refer to as clades (Figure 4-21). Clade-I comprised 24 of the 39 BED-NLRs (*Populus trichocarpa* and *Salix purpurea*) with only six BED-proteins that all derived from *Populus trichocarpa*. All carried only on BED domain, except for one BED-LRR protein (Appendix 8-18). This expansion of BED-NLRs is similar to what we observed in the Pooideae and Panicoideae.

We identified seven BED-NLRs and nine BED-proteins in Clade-II, originating from both *Populus trichocarpa* and *Salix purpurea* (Figure 4-21). Three of these proteins displayed Dimer_Tnp_hAT, DUF and double-BED domain architecture, one carried one Dimer_Tnp_hAT and one DUF domains, two only encoded DUF domains and the last two were single BED proteins (Appendix 8-18).

Clade-III is composed of two sub-clades (Figure 4-21). However, the Dimer_Tnp_hAT and DUF architecture was found in both sub-clades (Appendix 8-18). Clade-IV also comprises smaller divergent clades (Figure 4-21). We identified a total of 10 BED-proteins clustering with four BED-NLRs. Five of these BED-proteins showed a Dimer_Tnp_hAT and DUF domain composition, three encoded DUF only and the last two were single-BED proteins (Appendix 8-18).

Overall, we observed the same domain composition in BED-proteins whose BED domains cluster with BED-domains from BED-NLR proteins.

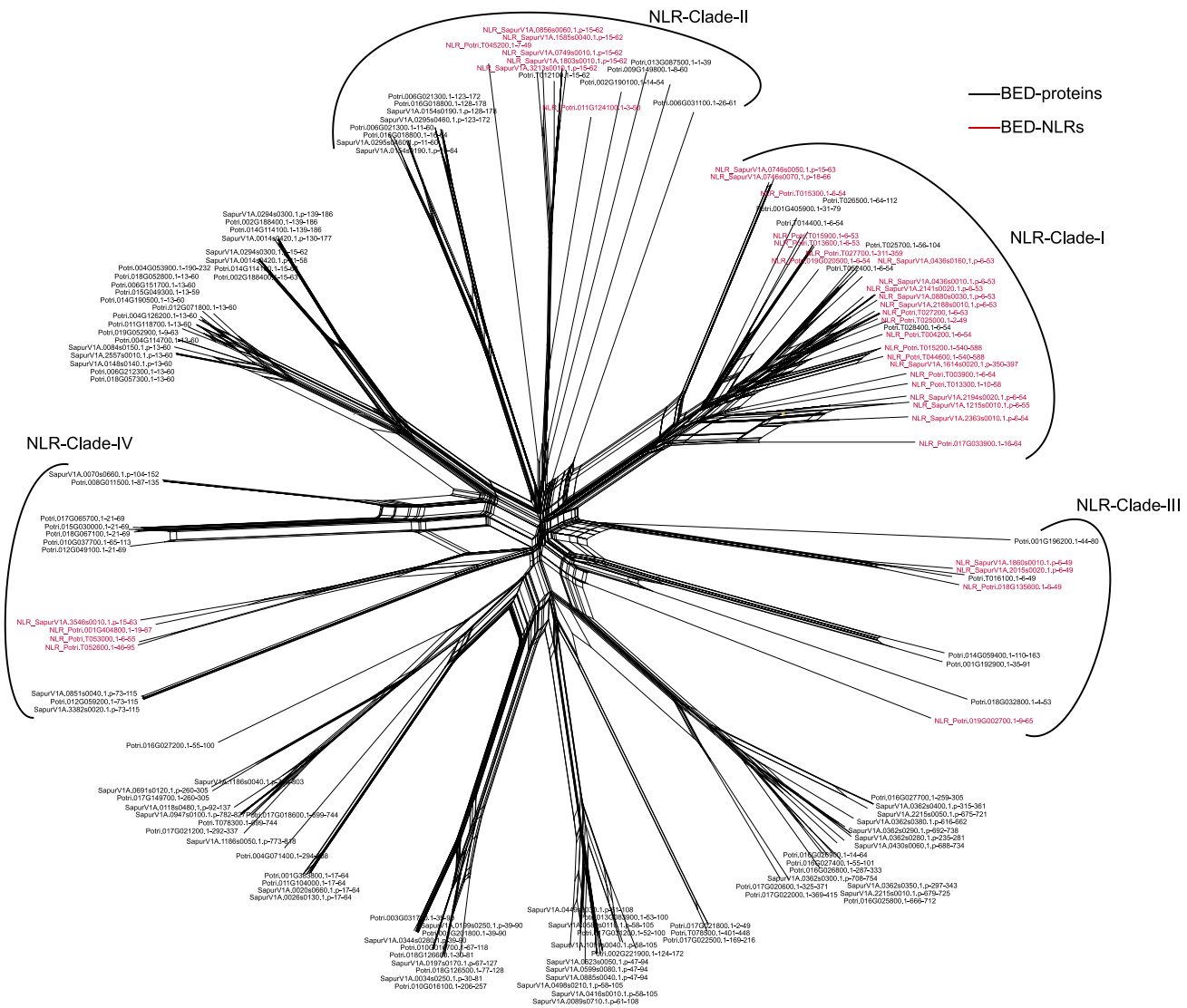


Figure 4-21. Neighbour-net analysis based on uncorrected P distances obtained from alignment of 152 BED domains including 39 BED- NLRs from the green group defined on Table 4-14 (Malpighiales).

BED-NLR are shown in red and other BED-containing proteins in black.

Summary

From this comparative study on BED domains derived from plants having BED-NLRs and BED-proteins in their genomes, we confirmed that:

- BED domains are highly divergent. This is demonstrated by the star-like shape of all networks we showed
- BED domains from BED-NLRs do not cluster altogether. Indeed, we identified more than one clade of BED-NLRs/BED-proteins in each network. Moreover, the different clades did not correspond to the different species investigated in one network. This might indicate that BED-NLRs emerged independently several times in plants.
- Some clades mostly contained single-BED domain proteins clustering with BED-NLRs. It would be interesting to determine whether the genomic positions of such proteins are close to BED-NLRs. Indeed, these single domain proteins could be mis-annotations.
- Overall, Dimer_Tnp_hAT was found in BED-proteins that clustered with BED-NLRs in all network, as well as proteins encoding DUFs only.

Is Dimer_Tnp_hAT significantly enriched in BED-proteins that cluster with BED-NLRs or do we observe a high amount of this additional domain because of its representation in BED-proteins in general? To answer this, we carried out one-way Fisher's exact tests for all additional domains we identified in BED proteins that clustered with BED-NLRs. This test allowed us to determine whether the proportion of a given domain in BED proteins that clustered with BED-NLRs was different of its proportion in BED-proteins in general.

Table 4-15. List of domains we found in BED-proteins that clustered with BED-NLRs in our split-network analyses (Figure 4-17, Figure 4-18, Figure 4-19, Figure 4-20, Figure 4-21) with associated Fisher's exact test to determine whether the proportion of a given domain in BED-proteins clustering in clade with BED-NLRs was greater than the proportion of this domain in BED-proteins in general (alternative hypothesis = greater). We showed in green *p-values* < 0.01 and in red *p-values* > 0.01

Domain	In clade	In all BED proteins	Fisher Exact test (p-value, alternative hypothesis = greater)
Dimer_Tnp_hAT	92	*339*	1.36E-14
DUF4413	44	*211*	1.96E-04
DUF659	71	*191*	2.20E-16
NAM	4	*21*	0.256
F-box	6	*8*	7.71E-05
Ank	1	*5*	0.4831
CENP-B_dimeris	2	*2*	0.01523
DnaJ	2	*2*	0.01523
Myb_DNA-bind	1	*2*	0.2319
Nop14	2	*2*	0.01523
DBD_Tnp_Hermes	1	*1*	0.1236
DUF295	1	*1*	0.1236
DUF1342	1	*1*	0.1236
DUF3420	1	*1*	0.1236
FAM177	1	*1*	0.1236
NPR1_like_C	1	*1*	0.1236
Sec34	1	*1*	0.1236
SWIB	1	*1*	0.1236
Tmemb_14	1	*1*	0.1236
single-BED	135	*1659*	1*
Total	363	2575†	

**p-value* < 2.2E-16 if tested with alternative hypothesis = less
†Includes the 124 BED-NLRs found in the 20 investigated proteomes

There were four domains that were significantly (*p-value* < 0.01) over-represented in BED-proteins clustering with BED-NLRs compared to their proportion in BED-proteins: Dimer_Tnp_hAT, DUF, and F-box domains. BED domains from Dimer_Tnp_hAT and F-box proteins are thus more similar to BED domains from BED-NLRs in plants in general. Additionally, single-BED proteins were significantly (*p-value* < 2.2E-16) under-represented in BED-proteins that clustered with BED-NLRs when compared to their proportion among BED-proteins in general.

In this Chapter, we used comparative genomics and split-network analyses to identify additional features associated with the BED-NLR domain organisation. This allowed us to generate new hypotheses regarding the role of the BED domains in BED-NLRs. We will further explore these hypotheses in Chapter 5.

4.4. Discussion

4.4.1. We identified *Yr7* and *Yr5* alleles in the wheat pangenome

Exploring nine additional wheat genomes allowed us to identify a new *Yr7* allele that was absent from the varieties we investigated in Chapter 3. Additionally, we found more occurrences of Claire-Yr5 and Cadenza-Yr5 in the wheat pangenome, although these alleles were located on Chromosome 2D.

The *Yr7* allele identified in Landmark, Mace and Stanley (Group 3) has only one SNP with Cadenza-Yr7. Moreover, the markers we designed in Chapter 3 are not able to differentiate between Cadenza-Yr7 and Landmark/Mace/Stanley-Yr7. This illustrate how crucial it is to have access to as much sequence information as possible across varieties to allow strict discrimination between alleles when designing markers. However, it is unknown whether the Landmark/Mace/Stanley-Yr7 allele is functional. Answering this will determine whether the *Yr7* KASP markers designed in Chapter 3 still can select functional *Yr7* alleles.

4.4.2. The *Yr* locus is conserved in wheat and wheat-related species

4.4.2.1. Phylogenetic relationship between NLRs located in the Yr locus

The syntenic region encompassing the *Yr* locus was overall conserved across the ten wheat assemblies and wheat related-species we investigated (Figure 4-3, Figure 4-10). The presence of two different regions containing NLRs with one cluster containing canonical NLRs only and one wider region containing BED-NLRs, could be traced back to rice (Figure 4-10). This was confirmed in a phylogenetic analysis on the NB-ARC domain (Figure 4-12), where the non-canonical NLRs were all more phylogenetically related among them than to the BED-NLRs. This suggest that the distinction between the canonical and BED-NLR predates the wheat/rice divergence, with loss of the BED-NLRs in the barley accession (Morex) we investigated.

We identified two main types of BED domains (BED-I and BED-II) in BED-NLRs located in the *Yr* syntenic region and this distinction predated the wheat/*Brachypodium distachyon* divergence (Figure 4-10). Additionally, the *Yr7* and *Yr5* clade only contained bread wheat, emmer wheat and goat grass BED-NLRs (Figure 4-12). This could either suggest that the evolutionary constraints applied to the NB-ARC and BED domains are different, or that BED-I-BED-II was the ancestral substructure and then BED-NLRs underwent differential losses of BED-I/BED-II with several occurrences of the BED-I-BED-II structure remaining. Alternatively, BED-I might have been the ancestral state, followed by the introduction of BED-II and differential loss of BED-I/BED-II in more recent BED-NLRs, again with several occurrences of the BED-I-BED-II structure remaining.

4.4.2.2. A subset of the BED-NLRs carries an NLS in the vicinity of the BED domain

Nuclear localisation signals (NLS) were identified in *Xa1* and the candidate gene for *Xo1*, both rice BED-NLRs[226]. We thus asked the question whether the presence of an NLS was a BED-NLR feature. We identified NLS in 14 of the 32 BED-NLR located in the *Yr* locus in bread wheat, emmer wheat, goat grass, *Brachypodium* and rice. The NLS were preferentially located ~30-40 residues downstream of the BED domain in BED-I and either ~30-40 residues downstream or ~10 residues upstream the BED domain in BED-II NLRs (Figure 4-13). In *Xa1* and rice BED-NLRs located in the *Yr* region, there were two predicted NLS flanking the BED domain (Figure 4-13). The NLS upstream the BED domain was similar to the corresponding region in BED-II NLRs.

Nuclear localisation of certain NLRs is required for resistance. For example, the nuclear localisation of barley Mla10 and *Arabidopsis* RPS4 is required for the expression of this resistance[269,270]. Furthermore, in the nucleus the activated Mla10 prevented the interaction between HvMYB6 transcription factor (positive regulator of immunity) with HvWRKY1 transcription factor (negative regulator of immunity) and thus allowed HvMYB6 to bind to its corresponding cis-elements[271]. The ensuing transcriptional reprogramming was necessary for the expression of Mla10-mediated resistance.

Based on the studies discussed above, nuclear localisation seems to be a requirement for the resistance mediated by certain NLRs. Whether the NLS identified in *Yr7* and *Yr5* are functional remains to be determined. We will explore this in Chapter 5.

4.4.3. An NLR cluster was identified in wheat varieties carrying a *Yr7* allele

NLRs tend to be organised into cluster in plant genomes (concept first described by Michelmore and Meyers (1998)[186]). The first definition of gene cluster was proposed by Holub (2001)[272] as follow: a gene cluster is a region in which two neighbouring homologous genes are < 200 kb apart. However, this was defined for *Arabidopsis thaliana*, which has a gene density of 15-32 open reading frames (ORFs) per 100 kb[273]. In the same study, the authors compared the gene organisation in *Gramineae* (maize, rice and barley) and in *Arabidopsis* and concluded that in *Gramineae* the coding region were grouped in large clusters interspaced with wide intergenic regions, whereas the gene distribution in *Arabidopsis* was fairly homogeneous across its genome[273].

In wheat the gene density tends to increase from the centromere to the telomeric regions[112,274]. A similar trend was observed in large plant genomes such as soybean[275] and maize[276]. NLR loci also tend to be located near the distal ends of chromosomes in wheat[195] and other species such as the Solanaceae[277], *Setaria italica*[278] and cotton[279]. Whether this is a direct implication of the gene distribution or a favoured distribution of NLR loci, however, remains to be tested. From the detailed analysis of each wheat chromosome, we observed that the gene density encompassing the *Yr* locus (from 600 Mb to 700 Mb) varied from 50 to 100 genes/Mb, or 5 to 10 genes per 100kb[112]. This is close to what was observed for *Arabidopsis*. Consequently, we can define the six uninterrupted NLRs genes, including the *Yr7* allele, spanning a 455 kb region in Landmark, Mace and Stanley as an NLR cluster. SY-Mattis carried five NLR genes in this cluster, lacking the *Yr7* allele.

Interestingly, this whole cluster seems to be absent from the other varieties included in this study (Figure 4-3). Indeed, this region did not contain any genes in Chinese Spring, Arina, Norin61 or Jagger and contained only one NLR in Lancer and Julius (nlr_4; Figure 4-3). Such intraspecies variation in NLRs contents is known in other plant species[277,280,281].

Furthermore, an NLR locus was located ~ 4 kb downstream of the *Yr7* functional allele. A homolog to this NLR was also present in Cadenza (98.4 % sequence identity), 5 kb downstream of *Yr7*. After correcting its sequence with RenSeq data and Sanger sequencing, we attempted to define the gene structure of this NLR locus. However, we could not define a continuous coding region in this locus. Could this NLR locus be an expressed pseudogene? About 7.2 % barley pseudogenes were found to be expressed[282] and a study found that 28 % of putative wheat pseudogenes on chromosome 3B were expressed[283]. It is unknown whether expressed pseudogenes have a role in plants. However, several studies in human and mammals showed that expressed pseudogenes have a role in regulating the transcription of their 'parent gene' (their protein-coding homolog) through the following processes: (i) gene expression suppression by natural antisense RNA; (ii) RNA interference by producing short interfering RNAs (siRNAs) and (iii) act as microRNA decoys (miRNA) (reviewed in Sen and Ghosh (2013)[284] and Pink et al., (2011)[285]). In this Chapter we only had access to RNA-seq data from Cadenza leaf samples under no induced biotic or abiotic stress. Based on this evidence only and the position of the NLR locus in close proximity to *Yr7*, we lack evidence to determine whether this locus has any function in *Yr7*-mediated resistance. Furthermore, this potential pseudogene only shared ~ 79 % with *Yr7* so it is unsure whether it could have a regulatory role via the mechanisms described above.

We did not find evidence of an NLR partner in the vicinity of *Yr7*. However, this does not invalidate the hypothesis that an additional component may be required in Yr7-mediated resistance. Indeed, NLRs recognising the pathogen (sensor) sometimes require the presence of another NLR to signal defense response (helper) that is not necessarily in close proximity in the genome[286]. For example, a major clade of NLRs in Solanaceae plant species forms a complex immunoreceptor network including multiple *NRC* (NLR-REQUIRED FOR CELL DEATH) helper NLRs that are required by numerous sensor NLRs involved in resistance against multiple pathogens[287]. Several sensors also rely on the same helper for defense signalling. For example, *NRC2*, *NRC3*, and *NRC4* helpers redundantly contribute to the immunity mediated by other sensor NLRs, including *Rx*, *Bs2*, *R8*, and *Sw5*[287]. Thus, finding only one candidate carrying mutations in *Yr7* in all *Yr7*-loss of function mutants does not invalidate the hypothesis of a required helper whose function is redundant. Further investigating the susceptible mutants in the Lemhi-Yr5 background whose susceptible phenotype is complemented in F_2 progenies derived from a cross between these lines and Vuka, which does not contain *Yr5*, may be a good starting point to address this hypothesis.

4.4.4. Chromosome-scale assemblies enabled comparison of the full *Yr* locus across ten varieties

NLR loci are among the most variable loci in different varieties of the same species. To address whether the conservation within the *Yr* locus was similar to that observed in genomic regions flanking it, we expanded the locus by ~ 8 Mb on each side and investigated gene content and whole genome sequence conservation across wheat varieties (Figure 4-5 to Figure 4-9).

4.4.4.1. Degree of conservation between the Yr locus and its flanking regions is variable across the different wheat groups

We defined three sub-groups on Figure 4-3 based on the sequence similarity of NLRs located in the *Yr* region and the whole architecture of the locus. These subgroups were conserved when we expanded the analysis to all genes and whole genomic region in +/- 8Mb flanking the *Yr* region. However, the degree of conservation between the *Yr* locus and its flanking region varied between the groups (Appendix 8-11). Indeed, the *Yr* locus had a higher SNP density than the flanking region in the alignment between Julius and Lancer, whereas we observed the contrary in the alignment between Landmark, Mace, Stanley and SY-Mattis (Appendix 8-11). Additionally, the SNP density in the *Yr* locus and its flanking region was comparable in the alignments between Arina, Chinese Spring, Jagger and Norin61.

Apart from Chinese Spring, all nine varieties are current elite cultivars. Furthermore, a *Yr7* allele was identified in Landmark, Mace and Stanley. There was comparable number of SNPs between Landmark/Stanley and Julius/Lancer in the whole region (266 and 348, respectively), whereas there was much lower number of SNPs in the *Yr* locus between Landmark/Stanley than Julius/Lancer. This suggests that the *Yr* locus might have evolved differently between these varieties. It is also tempting to speculate that given that the *Yr7* allele could be functional and/or *Yr7* might be linked to other traits of agronomic value (discussed in Chapter 3), selecting for this locus through breeding in Landmark/Stanley might have selected for less variant haplotypes than in Julius/Lancer. However, we cannot make strong conclusions here because we do not know the relationship between these varieties. Likewise, we would need to know what known locus they have been selected for and compare with other genomic regions that were not under breeding selection. It is nonetheless an interesting question to ask.

4.4.4.2. Structural re-arrangements between varieties observed in the Yr locus might be due to assembly errors

We observed small and large-scale structural re-arrangements in the *Yr* locus between wheat varieties. This included inversions, translocations and insertions/deletions (Figure 4-7, Figure 4-9). However, when looking at the alignments more closely it appeared that at least one of the breakpoints of the alignment coincided with a region containing 'Ns' in at least one of the investigated assemblies. We thus cannot discriminate between re-arrangements due to assembly errors and real structural re-arrangements. This has been already observed when comparing chromosome 2D between two wheat varieties[288]. Among 26 InDels larger than 100 kb that were identified between the two chromosomes, the authors discarded 22 based on the presence of Ns at two of the breakpoints and one of the four that were further analysed has Ns at one of the breakpoints. More focused sequencing effort, e.g. using BioNano Genomics, will be required to decipher whether the re-arrangements we observed are real or due to assembly errors.

4.4.5. Neighbour-net analyses allowed identification of a certain BED-containing protein families whose BED domain is similar to that of BED-NLRs in plants

We investigated 69 plant proteomes spanning the plant kingdom to determine the relative frequency of BED domains and BED domains integrated into NLR proteins. We found 66 out of 69 proteomes contained BED-containing proteins, of which only 20 contained BED-NLRs. The presence of BED-NLRs was found in separate clades in both monocots and dicots. This might suggest that either the presence of this domain in NLRs might

have occurred independently several times through plant evolution or it was lost in most of the plant genomes and only retained in a small subset.

The BED domain is short and highly variable (Figure 4-14). This renders phylogenetic analyses challenging. We used a phylogenetic method similar to the one we selected for the NB-ARC (Figure 4-12) to determine the relationship between BED domains in several plant species. However, the bootstrap value supporting the nodes were very low (data not shown). This can be due to recombination, hybridization, gene conversion, and gene transfer, which are evolutionary histories that are difficult to model with a tree[264]. To overcome this, we carried out neighbour-net analyses, which focuses on sequence similarities based on a distance matrix without inferring phylogenetic algorithms. Additionally, we split the proteomes into groups of phylogenetically close species in the attempt to reduce the noise produced by the high sequence variability across BED domains (Table 4-14). We thus cannot make hypothesis regarding the ancestral state of the BED domain in plants, but we can investigate which BED domains are close to each other based on sequence similarity.

We selected clusters containing both BED domains from BED-NLRs and BED domains from BED-containing proteins in each of the five neighbour-net analyses (Figure 4-17 to Figure 4-21). We then determined whether there was a certain BED-containing protein architecture that was significantly associated with BED-NLRs. Taking together results from the five networks, we found that the homodimerization domain of hAT transposases, several domains of unknown function (DUFs) and F-box domains were found more often in BED-proteins which clustered with BED-NLRs. This enrichment was significative when compared to the proportion of these additional domains found in BED-proteins in general (Table 4-15, Fisher-s exact test, *p-value* < 1.96E-4). However,

very few BED-F box proteins were identified in total (6 in clades with BED-NLRs and 8 in all BED-proteins) compared to homodimerization domain of hAT transposases (92 in clades and 339 in all BED-proteins). Furthermore, each of the five phylogenetic groups we analysed had BED domains from BED-NLRs clustering with BED-domains from BED-hAT proteins, whereas F-box proteins were only found in two groups. Providing the mode of action of BED-NLRs in conserved across plants and involves processes similar to what is proposes in the integrated decoy model, BED-hAT proteins have BED domains that are more similar to BED-NLRs than any other BED-proteins.

We discussed in Chapter 3 that in plants, the daysleeper proteins both carry a BED domain able to bind DNA and had deleterious effects on plant development in knock-out mutants[236]. Additionally, this transcription factor family also carries the homodimerization domain of hAT transposases and thus have been hypothesized to have arisen from neo-functionalization of a domesticated hAT transposase[237]. It is tempting to speculate that the BED domain in *Yr7* and *Yr5* could function in a similar way to WRKY domain in RRS1-R[45,231] (discussed in section 4.1.1.3), where the effector PopP2 acetylates both WRKY domains from transcription factors and RRS1-R and this prevent DNA binding. Furthermore, despite having a WRKY domain that is identical to RRS1-R, RRS1-S is not able to trigger defense response in presence of PopP2. The authors thus concluded that the extra residues at the C terminus of RRS1-R might have a role in activation of the defense response in the presence of PopP2. Similar reasoning could explain the difference in resistance spectra against *Pst* between *Yr5* and *YrSP*, despite having identical BED domains. Alternatively, BED domain is not directly involved in pathogen recognition.

Could some *Pst* effectors be able to trigger BED domains from the *daylsleeper* transcription factor family to facilitate infection? BED-NLR may thus play the role of effector traps and evolved a mechanism that allows detection of the modification of their own similar BED domain to trigger defense responses. Further experimental work will be required to validate this hypothesis and based on our current results, we still cannot exclude that BED domain in BED-NLRs could be involved in signalling.

4.4.6. Summary

Combining comparative genomics and neighbour-network analyses allowed us to uncover new features of functional BED-NLRs:

- presence of a Nuclear Localization Signal in the vicinity of the BED domain
- absence of a protein-coding NLR locus in head-to-head orientation with *Yr7*
- similarity between BED domains from BED-NLRs and BED domains from BED-hAT proteins.

These findings allowed us to refine our hypothesis regarding a potential mode of action of BED-NLRs in the frame of the integrated decoy model. We will further investigate the mode of action of *Yr7* in Chapter 5.

5. Functional characterisation of *Yr7*

5.1. Introduction

We showed in Chapter 4 that combining comparative genomics and neighbour-net analyses allowed us to refine our hypothesis regarding the role of the BED domain in *Yr7*-mediated response. We hypothesized that the BED domain functions in a similar way to the WRKY domain of RRS1-R, which the PopP2 effector is able to directly bind and acetylate specific residues[45,231]. However, it might be that the C-terminus region of RRS1-R is involved in defense response activation, given that the susceptible allele RRS1-S interacts in a similar way with PopP2 but does not trigger cell death[45,231]. PopP2 is also able to acetylate WRKY domains from WRKY transcription factors and this activity disables their ability to bind DNA[45,231].

In Chapter 4, we showed that certain BED domains from BED-containing proteins shared similarities with BED domains from BED-NLRs, especially BED-proteins having domain organisation similar to the transcription factor family *daysleeper* (BED-hAT homodimerization domain). We showed in Chapter 3 that only one residue was different between Yr7-BED and Yr5-BED and both Yr5-BED and YrSP-BED were identical, despite *Yr7*, *Yr5* and *YrSP* showing different resistance spectra to *Pst*. We thus hypothesized that other regions than the BED domain may be involved in isolate specificity in Yr7 and Yr5/YrSP, similarly to what we mentioned above with the C-terminus of RRS1-R likely to be involved in recognising PopP2, whereas RRS1-S does not.

We need further experimental evidence to support this hypothesis. We still cannot dismiss the possibility that the BED domain indirectly recognises the effector via guarding of a host protein. In Chapter 4, we gave the example of the Pii-2 NOI core motif

that is necessary for the interaction with host protein OsExo70-F3 upon effector binding of this target[248]. Alternatively, given BED domain conservation across BED-NLRs and that BED-NLRs tend to form their own clade in NLR phylogeny based on NB-ARC[88], it could be that this domain is important for defense response signalling.

In this section, we will use established methods to study the function of NLRs in plants and especially cereals, where heterologous systems may be required. This will provide an experimental system to test whether *Yr7* behaves in a similar way to characterised NLRs in cereals.

5.1.1. Transgenic approaches to validate resistance genes

We briefly discussed in Chapter 3 that the most common way to validate resistance genes in plants is to transform a susceptible variety with the gene of interest to determine whether it can provide resistance. This was successfully done for *Yr36*[63] and *Yr15*[146], among other cloned rust resistance genes. Additionally, *Sr22, Sr33, Sr35, Sr50* were also validated via transfer into a susceptible wheat cultivar (Fielder)[137,172,289,290].

However, although transgenic complementation provides a very strong evidence for the candidate gene to be the causal gene and validates the function of the transferred resistance gene, a negative result is not necessarily conclusive. Indeed, in rice it has been shown that the expression of the resistant phenotype upon gene transfer varied between rice varieties[185]. This can be due to the absence of other components involved in the expression of the resistance in the transformed cultivar or to negative interactions between the newly introduced gene and the genomic background of the transformed cultivar. For example, several studies reported *R*-gene suppression in wheat when introducing a new allele in a variety. For example, pairwise combinations of different

Pm3 alleles in F_1 hybrids and stacked transgenic wheat lines can result in suppression of *Pm3*-based resistance[291].

Additionally, F_1 hybrid necrosis occurs when crossing different strains of *Arabidopsis* thaliana (2 % of tested crosses)[292]. In one case, hybrid necrosis was due to epistatic interactions between two alleles from two different NLR loci, *DM1* and *DM2*, in the F_1[292]. Auto-necrosis is a phenotype milder than hybrid necrosis and it was also linked to the mechanisms of disease resistance[293]. In this case, the *Cf-2* allele from *Solanum pimpinellifolium* conferring resistance against *Cladosporum fulvum* caused auto-necrosis when transferred into *Solanum lycopersicum* (domesticated tomato)[293]. This auto-necrosis phenotype develops only when *S. lycopersicum* is homozygous for an allele at an independent locus[293]. Although *Cf-2* does not encode an NLR, this example shows that transformation of a susceptible variety/species with a resistance gene can sometimes lead to an auto-immune response.

On the other hand, in the case of NLR pairs, both partners are required for the expression of resistance. For example, it has been shown that neither *RGA4* nor *RGA5* alone were able to confer resistance in rice [232]. However, when both NLRs were transferred into this variety, it recapitulated *RGA4*/*RGA5*-mediated resistance against *Magnaporthe oryzae*[232]. Thus, if the recipient of the transgene lacks other partners involved in the expression of the resistance, a susceptible phenotype is likely to be observed in the transgenic plants. This phenotype cannot be interpreted, however, as the gene having no effect on disease resistance response.

We provide evidence in Chapter 3 that we cloned *Yr7* and *Yr5/YrSP*. Although transgenic complementation might not be conclusive, its outcome will still allow us to understand

better how *Yr7* and *Yr5* work. Indeed, if the susceptible variety expresses the corresponding resistance upon transformation, then it provides more evidence that *Yr7* and *Yr5* work as singletons (i.e. not in a pair) or with partners that are conserved in wheat, similarly to what we discussed in Chapter 4 with the NRC network in Solanacaea[287]. Alternatively, if *Yr7* and *Yr5* transgenic plants do not express the corresponding resistance, it might be that the recipient variety is lacking an interacting partner or is not compatible with *Yr7* or *Yr5*.

5.1.2. *Nicotiana benthamiana* as a heterologous system to study NLR function in plants

N. benthamiana is a well-established heterologous system to study plant-pathogen interactions[294]. There are numerous examples in the literature reporting its use in transient expression assays to study NLR function including Hypersensitive Response (HR) signalling, NLR/effector interaction, cellular localisation, etc. This includes NLR derived from monocots such as stem rust resistance genes introgressed into wheat *Sr33*[172,295](*Aegilops tauschii*), *Sr35*[296] (*Triticum monococcum*), *Sr50*[297] (rye), well-characterised rice NLR pairs *RGA4*/*RGA4*[232,239] and *Pik-1*/*Pik-2*[234,298] or barley *Mla10*[215,269], among many others. Hypotheses regarding the function of NLRs derived from monocot plant species are thus widely tested in the dicot *N. benthamiana*.

5.1.2.1. Recapitulating NLR signalling in N. benthamiana

HR signalling is the main read-out used to study NLR function in *N. benthamiana*. It has been reported that certain singleton NLRs, which can both sense and signal presence of the corresponding pathogen effectors, are able to trigger cell-death in *N. benthamiana* in the absence of the effector (reviewed in Adachi et al., 2019[299]). This includes *Sr50*[297]

(rye), *L6*[79] (flax), and several Arabidopsis NLRs (*RPP13*[300], *RPS5*[301] and *ZAR1*[302]). This suggest that these NLRs are repressed or rendered inactive in their host to prevent constitutive activation. Defence response signalling through these NLRs can thus be recapitulated in *N. benthamiana*.

HR signalling resulting from NLR/effector co-expression occurring in *N. benthamiana* is also widely used as a proof of interaction *in planta*. For example in wheat, transiently co-expressing *Sr50* and *AvrSr50* led to a stronger HR than the sole expression of *Sr50* in *N. benthamiana*[297]. Similar results were observed when transiently co-expression *Sr35* and *AvrSr35*[296]. The *AvrSr35* effector was further validated in the host (wheat) via purification and infiltration of the protein in susceptible and resistant wheat cultivars[296]. This provides further evidence supporting the suitability of using *N. benthamiana* to study NLR/effector interactions, even if the NLR is derived from monocot.

5.1.2.2. *Testing the ability of the NLR gene of interest to signal in N. benthamiana*

It is important to ensure that the gene of interest is actually able to signal in *N. benthamiana* if the sole transient expression of the NLR of interest does not trigger HR. Mutations in the MHD motif following the NB-ARC domain can lead to auto-activity in certain NLRs[303]. This was described for the *Solanum tuberosum* NLR *Rx*, conferring resistance against Potato virus X[304]. The authors demonstrated that a D to V mutation in the *Rx* MHD motif led to cell-death when transiently expressed in *N. benthamiana*. Further work linked this mutation in the MHD motif to favour ATP over ADP binding in flax gene *M*, leading to cell-death[305]. These mutations can thus be used to generate potential auto-active mutants and assess whether the NLR of interest is able to signal in *N. benthamiana*.

We hypothesized that *Yr7* does not belong to an NLR pair based on the fact that we did not find evidence of a partner in head to head orientation in close proximity to *Yr* locus (Chapter 3). To test this, we developed a transgenic approach in wheat to validate the function of *Yr7* and set-up transient expression assays in *N. benthamiana* to test whether *Yr7* is able to signal in this heterologous system. This will allow us to determine whether *N. benthamiana* is a suitable system to study *Yr7*-mediated resistance. We focussed on *Yr7* because as it was the first gene we cloned and was thus available for the experiments.

5.2. Materials and Methods

5.2.1. Developing transgenics for *Yr7*

A summary of the generation of the construct and wheat transformation is provided in Figure 5-1.

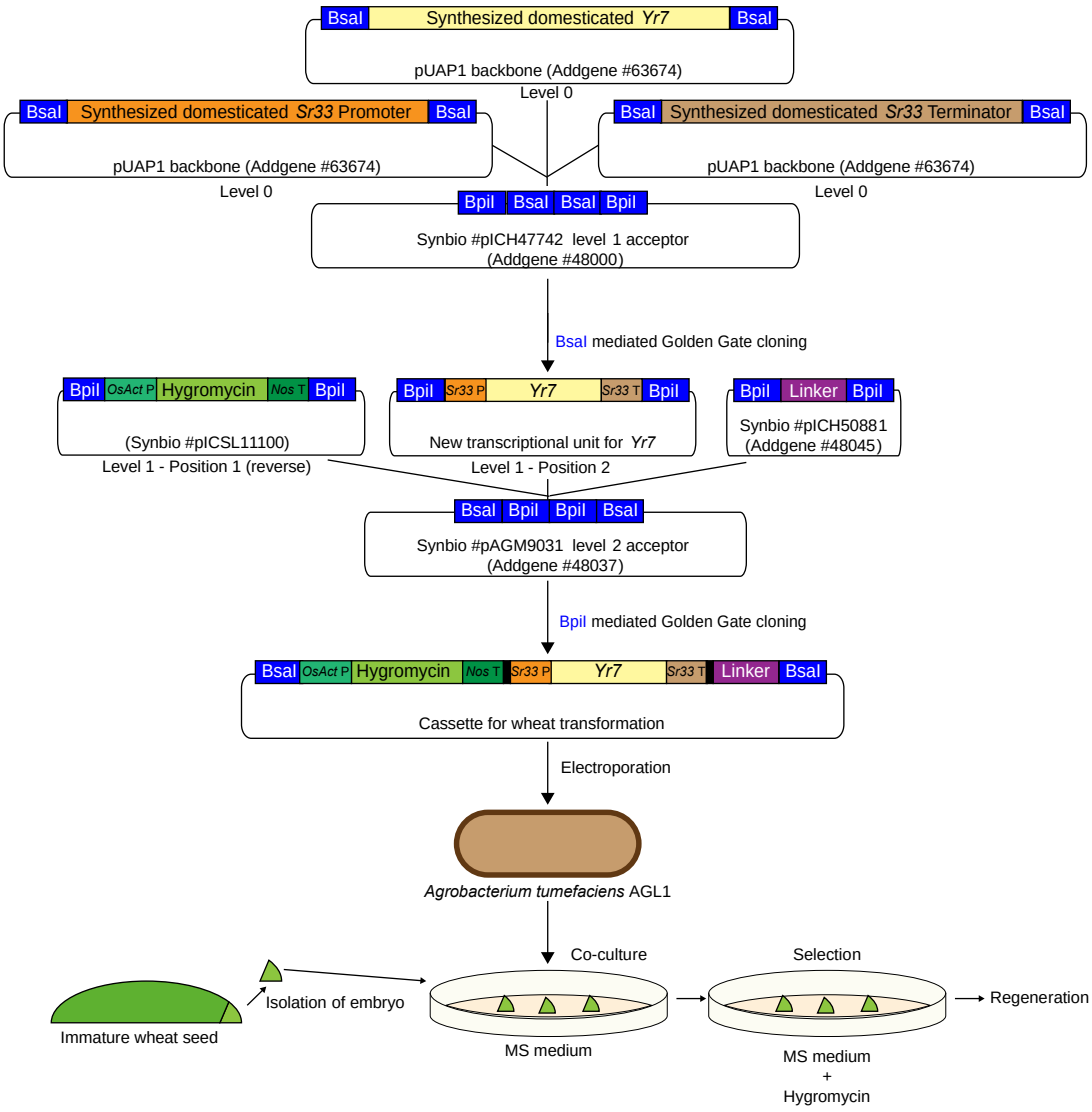


Figure 5-1. Summary of the cloning reactions and wheat transformation with *Yr7* cassette described in section 5.2.1.1

5.2.1.1. Generation of the Yr7 cassette for wheat transformation

We defined the full length of the *Yr7* locus in Chapter 3 (4,989 bp). We included both exons and intron in the construct. We used the Golden Gate approach[306] to generate the full cassette (Figure 5-1). As this cloning technique relies on *Bsa*I and *Bpi*I restriction enzymes, we first converted (i.e. domesticated) all five *Bsa*I and *Bpi*I restriction sites found in the *Yr7* sequences. We used the redundancy of the genetic code to not alter the derived protein product. We then synthesized the domesticated 4,989 bp *Yr7* product with flanking *Bpi*I sites to allow its ligation in pUAP1 level 0 acceptor (Addgene #63674).

We selected the regulatory elements (promoter, 2,381 bp and terminator, 1,405 bp) of the *Sr33* stem rust resistance gene[307]. We chose these as they were shown to work on the cultivar Fielder[307]. Both *Sr33* promoter and terminator were also cloned in pUAP1. We assembled this level 1 transcription unit (*Sr33P- Yr7 – Sr33T*) in the level 1 position 2 acceptor plasmid pICH47742[308] via *Bsa*I Golden Gate cloning (Figure 5-1). We then assembled this unit with a selection cassette including the rice *Act4* promoter, coding sequence of the hygromycin resistance and the Nos terminator (OsAct4P – Hygromycin - NosT) in reverse orientation (pICSL11100, http://synbio.tsl.ac.uk) and an end linker (pICH50881, Addgene #48045) to generate the level 2 cassette in pAMG8031 (Addgene #48037). We transformed the *Yr7* cassette into *Agrobacterium tumefaciens* AGL1 strain[309] via electroporation (Figure 5-1). We used Sanger sequencing to verify the sequence of the constructs at each step (Appendix 8-19)

5.2.1.2. Transformation of Fielder with Yr7 cassette

The constructs were subsequently introduced into *T. aestivum* cv. Fielder by Agrobacterium-mediated inoculation of 200 immature embryos (Figure 5-1). This part

was carried out by Sadiye Hayta (BRACT team, John Innes Centre) as described in Rey et al., 2018[310]. Briefly, after 3 days co-cultivation with *Agrobacterium*, immature embryos were selected on 15 mg/l hygromycin during callus induction for 2 weeks and 30 mg/l hygromycin for 3 weeks in the dark at 24°C on Murashige and Skoog medium[311] 30 g/l Maltose, 1.0 g/l Casein hydrolysate, 350 mg/l Myo-inositol, 690 mg/l Proline, 1.0 mg/lThiamine HCl[312] supplemented with 2 mg/l Picloram, 0.5 mg/l 2,4-Dichlorophenoxyacetic acid (2,4-D). Regeneration was under low light (140 µmol*m^{-2} *s^{-1}) conditions on MS medium with 0.5 mg/l Zeatin and 2.5 mg/l CuSO45H2O (Figure 5-1).

We obtained a total of 17 regenerated transgenic (T_0) plants. Confirmation of the presence of the transgene was done by PCR with the primers listed Appendix 8-19. Copy number variation was assessed by q-PCR by iDNA Genetics Norwich Research Park, UK (Table 5-1). Because the first batch of T_0 plants containing five lines, we first advanced these to T_1 and T_2 based on the presence of the transgene (see section below).

Table 5-1. Copy number variation in the Fielder+Yr7 T_0 lines.
The five lines highlighted in orange correspond to the lines we advanced to T_1 and T_2 generations. The two lines highlighted in red were kept as negative controls.

T_0 Line	Copy number
Yr7-1	2
Yr7-2	1
Yr7-3	1
Yr7-4	12
Yr7-5	6
Yr7-6	1
Yr7-7	1
Yr7-8	1
Yr7-9	19
Yr7-10	2
Yr7-11	3
Yr7-12	9
Yr7-13	2
Yr7-14	1
Yr7-15	10
Yr7-16	0
Yr7-17	0

5.2.1.3. Genotyping

We used the same primers as above to determine whether the T_1 and T_2 plants contained the *Yr7* transgene. We tested five to fifteen plants per line depending on how many T_1 plants were positive for the transgene in the first five. We performed a similar analysis in the T_2 plants.

5.2.2. PCR amplification and Golden Gate cloning of *Yr7* and its variants

A summary of the generation of the constructs and transient expression in *N. benthamiana* is provided in Figure 5-2.

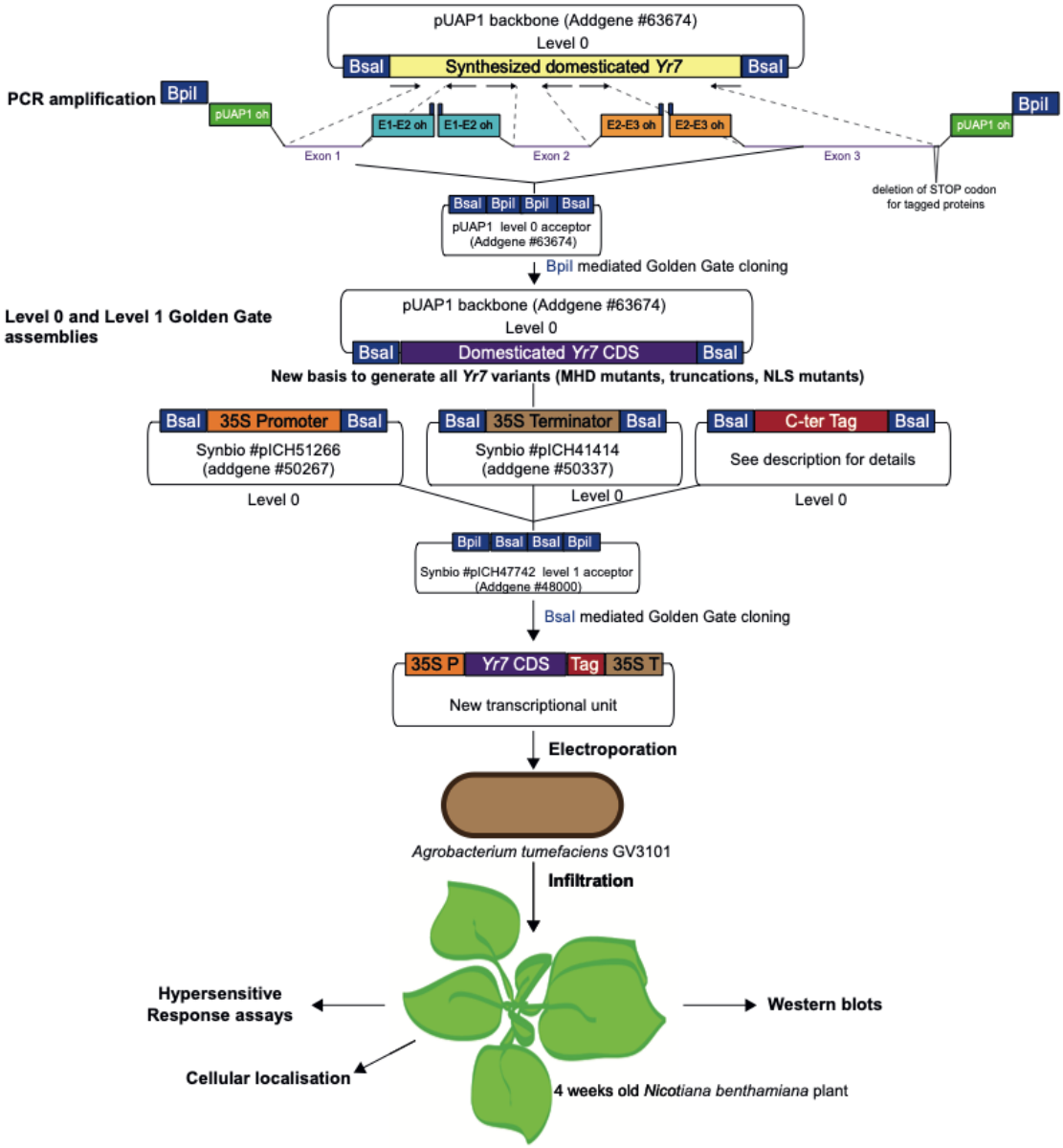


Figure 5-2. Summary of the generation of the different *Yr7* variants for transient expression in *N. benthamiana*

5.2.2.1. Yr7 coding region for N. benthamiana assays

We assembled the coding DNA sequence of *Yr7* (*Yr7* CDS) via amplifying the three exons by PCR with primers containing tails with corresponding overhangs and *Bpi*I restriction sited for *Bpi*I Golden Gate-mediated cloning in pUAP1 (Figure 5-2, Appendix 8-20). Two different constructs were generated: one with its STOP codon and one without to allow construction of tagged-Yr7. This basis was subsequently used as a template to generate mutations in the MHD motif of *Yr7* (D646V) and the truncations in *Yr7* (exon1, exon2 and AA201, Appendix 8-20, Appendix 8-21).

In parallel, we synthesized a codon-optimised version of *Yr7* CDS to generate two other truncations in *Yr7* (AA242 and AA308) and the mutations in the predicted NLS of *Yr7*. We hypothesized that the codon-optimised version would yield higher protein expression, but this was not the case (results not shown).

5.2.2.2. MHD mutants in Yr7

We discussed in the introduction how mutations in the MHD motif of certain NLRs led to auto-activity in *N. benthamiana* in section 5.1.2.2. We identified the MHD motif in *Yr7* and we tested whether a D to V mutation in the MHD motif would produce auto-activity in *N. benthamiana*. We used Golden Gate cloning with primers containing a mismatch to introduce the corresponding single base pair substitution leading to the D to V amino-acid change in the Yr7 sequence (primers listed in Appendix 8-20).

5.2.2.3. Yr7 truncations and mutants for cellular localization

We generated truncations in the *Yr7* CDS with the primers listed in Appendix 8-21. We obtained single exon transcriptional units: exon 1, exon 2 and exon 1 + exon 2 (referred to as AA201) lacking STOP codon in the 3’ end to allow for recombinant protein

generation with tags for western blots and cellular localisation (Appendix 8-22). We generated two other truncations from the start codon to the residue #242 and #308 (referred to as AA242 and AA308, respectively) that contain the predicted NLS to determine whether it is functional in *N. benthamiana.*

5.2.3. Transient assays in *N. benthamiana* (*Yr7*)

5.2.3.1. Infiltrations

Constructs described in Appendix 8-22 were transferred into *A. tumefaciens* GV3101 via electroporation. An overnight 10 mL liquid culture (LB medium + selection antibiotic) of transformed *A. tumefaciens* was prepared for infiltration in *N. benthamiana* as follow. We centrifuged the overnight liquid culture 5 mins at 4,000 rpm and resuspended the pellet with infiltration buffer (10 mM $MgCl_2$, 10mM MES, pH5.6, 150 µM acetosyringone). We performed this step three times before adjusting the $OD._{600}$ to the desired value before infiltration. We infiltrated 4-6 weeks old *N. benthamiana* leaves with a 1mL plastic syringe. When *Agrobacterium* clones expressing *P19* (silencing suppressor, pICSL11029 in http://synbio.tsl.ac.uk) were co-infiltrated with a construct, we diluted the $O.D._{600}$ of the *P19* culture to 0.1 prior infiltration.

For one given construct, we infiltrated two leaves per plant and carried out the experiment three independent times to validate protein expression. For protein expression, we harvested the tissue at different days post-infiltration to select the time where the best expression was recorded for further analyses. For cellular localization we harvested tissue at 1.5 dpi before preparing the samples for confocal microscopy.

For Hypersensitive Response assays (HR assays), we harvested leaves at least 7 dpi before scoring for necrosis as described in Maqbool et al., 2015[298]. We used Mla10-HA recombinant protein as a positive control for HR and Pikp2-HA as a negative control. Mla10-HA was provided by Hiroaki Adachi (Sainsbury Laboratory) and Pikp2-HA by Thorsten Langner (Sainsbury Laboratory).

5.2.3.2. Western blots

We grinded harvested leaf tissue in liquid nitrogen and added 1:2 weight:volume of protein extraction buffer (10% glycerol, 25mM Tris pH 7.5, 1mM EDTA, 150mM NaCl, 10 mM DTT, 2% w/v PVPP, protease inhibitor cocktail (Sigma #P9599), 0.15% NP-40). We centrifuged the samples 3,000 g for 10 min at 4 °C and filtered the supernatant in 2.5 mL syringes with 250um filters. We centrifuged the filtrate 13,000 rpm for 10 mins. We used 20uL to mix with 5X SDS loading buffer and reserved the remaining supernatant at -80 °C.

We loaded the protein samples in precast acrylamide gels (Mini-PROTEAN TGX, BioRad) and samples were run 20 mins at 70 V followed by 35-45 mins at 110 V in Mini-PROTEAN Tetra tanks (BioRad) with running buffer (25 mM Tris, 192mM Glycine, 0.1 % SDS, pH 8.3). Transfer onto nitrocellulose membrane was performed with Trans-Blot Turbo transfer system (BioRad) following the manufacturer instructions.

Membranes were incubated 2 hours with blocking buffer (1X TBS, 5 % milk) on a shaker. We washed the membranes three times 10 mins with 1X TBS + 0.20 % Tween20 and one last time 10 mins with 1X TBS. We probed the membranes 2 hours or overnight incubation at 4C with corresponding antibody (1:8,000 dilution in 1 X TBS + 5 % milk) in the case of anti-HA-HRP conjugated antibody, or 2 hours or overnight incubation with

anti-YFP antibody from rabbit (1:8,000 dilution in 1 X TBS + 5 % milk) followed by another 2 hours incubation with anti-rabbit-HRP antibody (1:10,000 dilution in 1 X TBS + 5 % milk). We washed the membranes as described in the first step before developing with Pierce™ ECL Western Blotting Substrate (ThermoFisher Scientific), following the manufacturer instructions. Chemiluminescence was capture with ImageQuant LAS-4000 (GE Life Sciences).

5.2.4. Cellular localization of *Yr7* truncation and NLS mutants

Transient expression in *N. benthamina* was performed as described in the section above with the *Yr7* truncations tagged with YFP. Infiltrated leaves were harvested two days after infiltration and kept on imbibed filter paper until observation under confocal microscope (Leica SP5). We used two plants per construct and two leaves per plant were harvested. Five fragments from each leaf were cut with a scalpel and mounted in water on microscope slides. Argon ion excitation laser (514 nm) was used to observe YFP-related fluorescence in the samples. Observations were done with x10 and x20 objectives.

5.3. Results

5.3.1. *Yr7* transgenics genotyping

We transformed bread wheat cultivar Fielder with *Yr7* driven by *Sr33* regulatory elements to determine whether *Yr7* alone is sufficient to express *Yr7*-mediated resistance in wheat. We regenerated 17 independent T_0 plants from which five were advanced to T_1 and T_2 generation based on genotyping. We tested between 6 and 18 T_1 plants per T_0 parent depending on how many T1 plants were found positive in the first batch of five plants (Table 5-2).

Table 5-2. Summary of the T1 plants genotyping for the presence of the *Yr7* transgene

T_0 line	**Transgene copy number in T_0**	**#Plants tested**	**#Positive for the transgene**
Yr7-1	2	6	6 (100 %)
Yr7-2	1	17	5 (29 %)
Yr7-3	1	18	7 (39 %)
Yr7-4	12	6	6 (100 %)
Yr7-5	6	6	6 (100 %)

All T_1 tested plants derived from T_0 containing > 1 copies of the transgene were positive. T_1 plants derived from T_0 containing one copy of the transgene were segregating for the presence of the transgene (Table 5-2). We recovered between 29 and 39 % of positive T_1 plants in the progeny. We kept four positive T_1 plants and two negative T_1 plants for Yr7-2 (Yr7-2-13, 14, 15, 21, 24 and Yr7-2-19, 23 respectively) and the seven positive T_1 plants for Yr7-3 (Yr7-3-13, 14, 15, 18, 19, 20, 21) to advance to T_2. We advanced all six positive T_1 plants for Yr7-1, Yr7-4 and Yr7-5 to T_2 generation.

We subsequently tested two to five T_2 plants per T_1 line to identify the ones carrying the *Yr7* transgene (Table 5-3). Most of the T_2 plants derived from T_1 plants that were positive for the presence of the transgene were also positive for the presence of the transgene (e.g Yr7-1 derived T_2 plants). All the T_2 plants derived from T_1 plants that were negative for

the presence of the transgene were also negative for the presence of the transgene (Yr7-2-19 and 23). Some lines were still segregating for the presence of the transgene due to the presence of both positive and negative progeny lines (Yr7-2-21, Yr7-3-15, 19 and 20, Yr7-5-16 and 15).

Based on these results, we selected two to three T_2 lines to advance to T_3 and test for *Yr7*-mediated resistance against *Pst*. We included both positive and negative lines (see 'x' symbol in Table 5-3). Tests were planned to be performed by Peng Zhang (University of Sydney) and seeds have been sent. Unfortunately, we did not have the results of the pathology tests at the time the thesis was written. Lines that were sent are summarised in Table 5-4. Both Fielder+Yr5 and Fielder+YrSP transgenics generated by Jianping Zhang (CSIRO) showed expression of the corresponding resistance (Jianping Zhang, personal communication). These transgenes were analogous to the ones we used for *Yr7* (*Yr5* and *YrSP* gDNA both under the control of Sr33 regulatory elements and additional construct were generated with the *Yr5* and *YrSP* putative regulatory elements).

Table 5-3. Summary of genotyping of the T_2 plants (Fielder + *Yr7*).
The third column show which T_2 plants will be advanced to T_3 generation for pathology tests to determine whether *Yr7* is functional in Fielder

Line	Presence of transgene	Selected for phenotyping
Yr7-1-1-1	+	x
Yr7-1-1-2	+	x
Yr7-1-1-3	+	
Yr7-1-1-4	+	
Yr7-1-1-5	+	
Yr7-1-2-1	+	
Yr7-1-3-1	+	
Yr7-1-3-2	+	
Yr7-1-4-3	+	
Yr7-1-4-4	+	
Yr7-1-6-1	+	
Yr7-2-13-1	+	x
Yr7-2-13-2	+	
Yr7-2-14-3	-	x
Yr7-2-14-4	-	
Yr7-2-15-1	+	x
Yr7-2-15-2	+	
Yr7-2-15-3	+	
Yr7-2-15-4	+	
Yr7-2-15-5	+	
Yr7-2-19-1	-	
Yr7-2-19-2	-	
Yr7-2-19-4	-	
Yr7-2-21-1	+	
Yr7-2-21-3	+	
Yr7-2-21-4		
Yr7-2-21-5	+	
Yr7-2-23-1	-	
Yr7-2-23-2	-	
Yr7-2-24-1	+	
Yr7-2-24-2	+	
Yr7-2-24-3	-	
Yr7-2-24-4	+	
Yr7-2-24-5	+	
Yr7-3-13-1	+	x
Yr7-3-13-2	+	
Yr7-3-14-1	+	
Yr7-3-14-2	+	
Yr7-3-15-1	-	
Yr7-3-15-3	+	
Yr7-3-18-1	+	
Yr7-3-18-2	+	
Yr7-3-19-1	+	
Yr7-3-19-2	-	
Yr7-3-20-1	-	x
Yr7-3-20-2	-	
Yr7-3-20-3	+	x
Yr7-3-20-4	+	
Yr7-3-21-1	+	
Yr7-3-21-2	+	
Yr7-4-1-1	+	x
Yr7-4-1-2	+	
Yr7-4-2-1	+	
Yr7-4-2-2	+	
Yr7-4-3-1	+	
Yr7-4-3-2	+	
Yr7-4-4-1	-	
Yr7-4-4-2	+	
Yr7-4-5-1	+	
Yr7-4-5-2	+	x
Yr7-4-6-1	+	
Yr7-4-6-2	+	x
Yr7-5-16-1	-	x
Yr7-5-16-2	+	x
Yr7-5-16-3	+	
Yr7-5-16-4	+	
Yr7-5-16-5	+	
Yr7-5-14-1	+	
Yr7-5-14-2	+	
Yr7-5-15-1	-	
Yr7-5-15-2	+	
Yr7-5-18-1	+	
Yr7-5-18-2	+	

Table 5-4. List of the Fielder-Yr7 transgenic lines that will be tested for the expression of Yr7 resistance by Peng Zhang (University of Sydney).
Lines shown in dark orange will be used as negative controls.

Line	Generation	Copy number in T_0 plants	Presence/absence transgene in T_2 plants
Yr7-1-1-1	T3 seeds	2	+
Yr7-1-1-2	T3 seeds	2	+
Yr7-2-13-1	T3 seeds	1	+
Yr7-3-13-1	T3 seeds	1	+
Yr7-2-14-3	T3 seeds	1	-
Yr7-5-16-1	T3 seeds	6	-
Yr7-5-16-2	T3 seeds	6	+
Yr7-2-15-1	T3 seeds	1	+
Yr7-3-20-1	T3 seeds	1	-
Yr7-3-20-3	T3 seeds	1	+
Yr7-6	T1 seeds	1	not tested
Yr7-7	T1 seeds	1	not tested
Yr7-8	T1 seeds	1	not tested
Yr7-9	T1 seeds	19	not tested
Yr7-10	T1 seeds	2	not tested
Yr7-11	T1 seeds	3	not tested
Yr7-12	T1 seeds	9	not tested
Yr7-13	T1 seeds	2	not tested
Yr7-14	T1 seeds	1	not tested
Yr7-15	T1 seeds	10	not tested
Yr7-16	T1 seeds	0	not tested
Yr7-17	T1 seeds	0	not tested

5.3.2. Optimisation of *Yr7* protein expression in *N. benthamiana*

To set-up a transient expression assay in *N. benthamiana*, we tested three different OD_{600} for the *A. tumefaciens* infiltrations (Figure 5-3). This will allow us to determine the best conditions to study *Yr7* and its variant in this heterologous system. Additionally, we conducted HR assays to determine whether *Yr7* is auto-active in *N. benthamiana*.

No expression was recorded for OD_{600} = 0.3 (data not shown). Protein expression of Yr7-HA was very similar at OD_{600} = 0.5 and 1 at both 1 and 2 dpi. However, no expression was recorded at 3 and 4 dpi. Pikp2-HA expression was higher than what we observed for

Yr7-HA. Expression of Yr7-HA did not trigger HR in *N. benthamiana*, whereas our positive control Mla10-HA and negative control Pikp2-HA produced their expected behaviour in *N. benthamiana* (HR and no HR, respectively). Yr7 was thus not auto-active in *N. benthamiana* in the tested conditions.

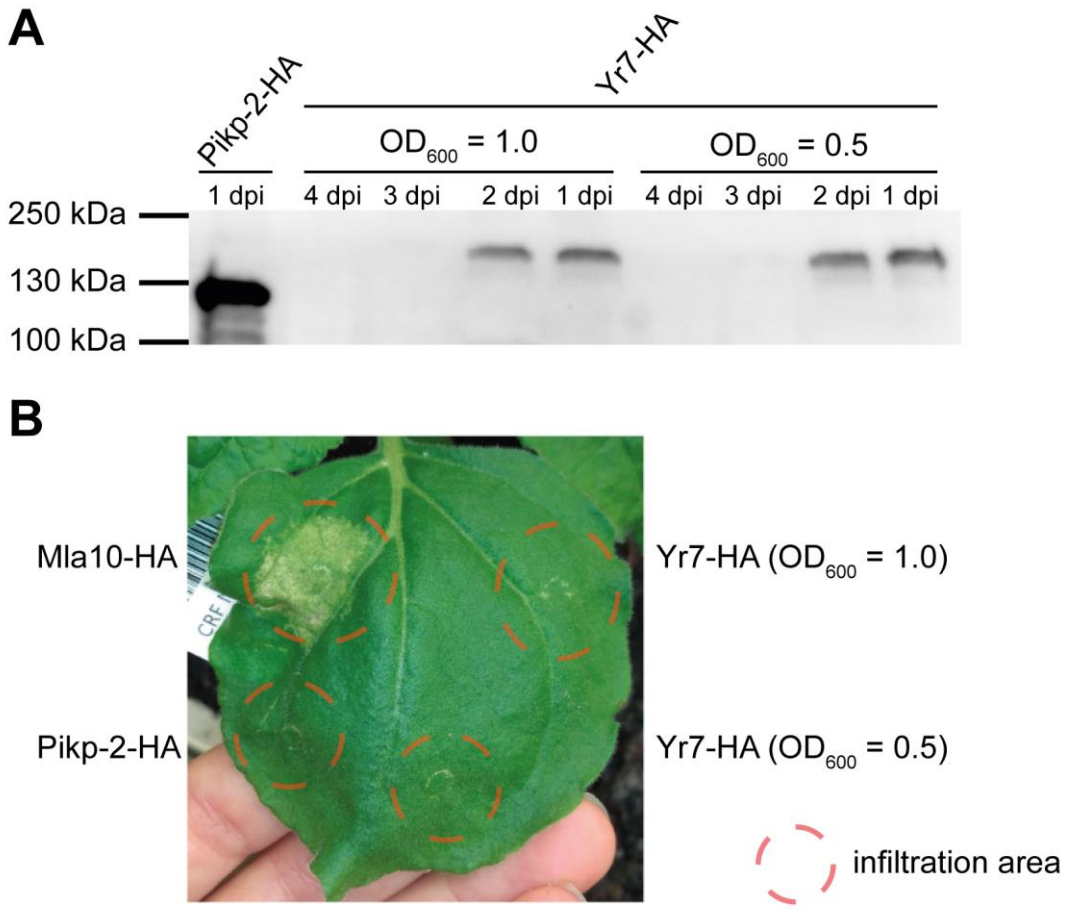


Figure 5-3. Optimisation of Yr7 protein expression in *N. benthamiana*.

A: Western blots showing the expression of Yr7-HA and Pikp2-HA 1, 2, 3- and 4-days post infiltration (dpi).

B: Hypersensitive Response assays (HR assays) recorded 7 dpi. *Mla10* was used as a positive control for HR and *Pikp2* construct as a negative control. Protein expression levels were also assessed for the Pikp2 negative control.

5.3.3. Hypersensitive Response assays with *Yr7* variants in *N. benthamiana*

5.3.3.1. Mutants in the MHD motif of Yr7

We tested whether Yr7 would behave in a similar manner to NLR singletons such as *Sr35* and *Mla10* when transiently expressed in *N. benthamiana* and induce a HR in Figure 5-3. Given that we did not observe any HR, we asked the question whether alteration of the MHD in Yr7 would induce auto-immunity, as observed in Rx^{304}. We thus induced D to V mutation in Yr7 MHD motif and transiently expressed these mutants in *N. benthamiana*.

The Pikp2-HA positive control showed high expression at 1 dpi (Figure 5-4A), as expected, and its expression was slightly higher in the presence of P19. Expression patterns of Yr7-D646V with or without P19 were different with Yr7-D646V expressed at 1 dpi only and Yr7-D646V + P19 showing an increased expression from 2 to 4 dpi (Figure 5-4A). It is unclear why Yr7-D646V was not expressed at 2 dpi, as shown in Figure 5-3. Overall, the D646V substitution did not alter the expression of Yr7.

Mla10 induced a strong HR in *N. benthamiana*, whereas Pikp2-HA or Pikp2-HA + P19 did not, as expected (Figure 5-4B). No HR was observed in the Yr7-D646V or Yr7-D646V + P19 infiltration areas (Figure 5-4B). Mutations in the MHD motif of Yr7 thus did not induce auto-activity in the tested conditions.

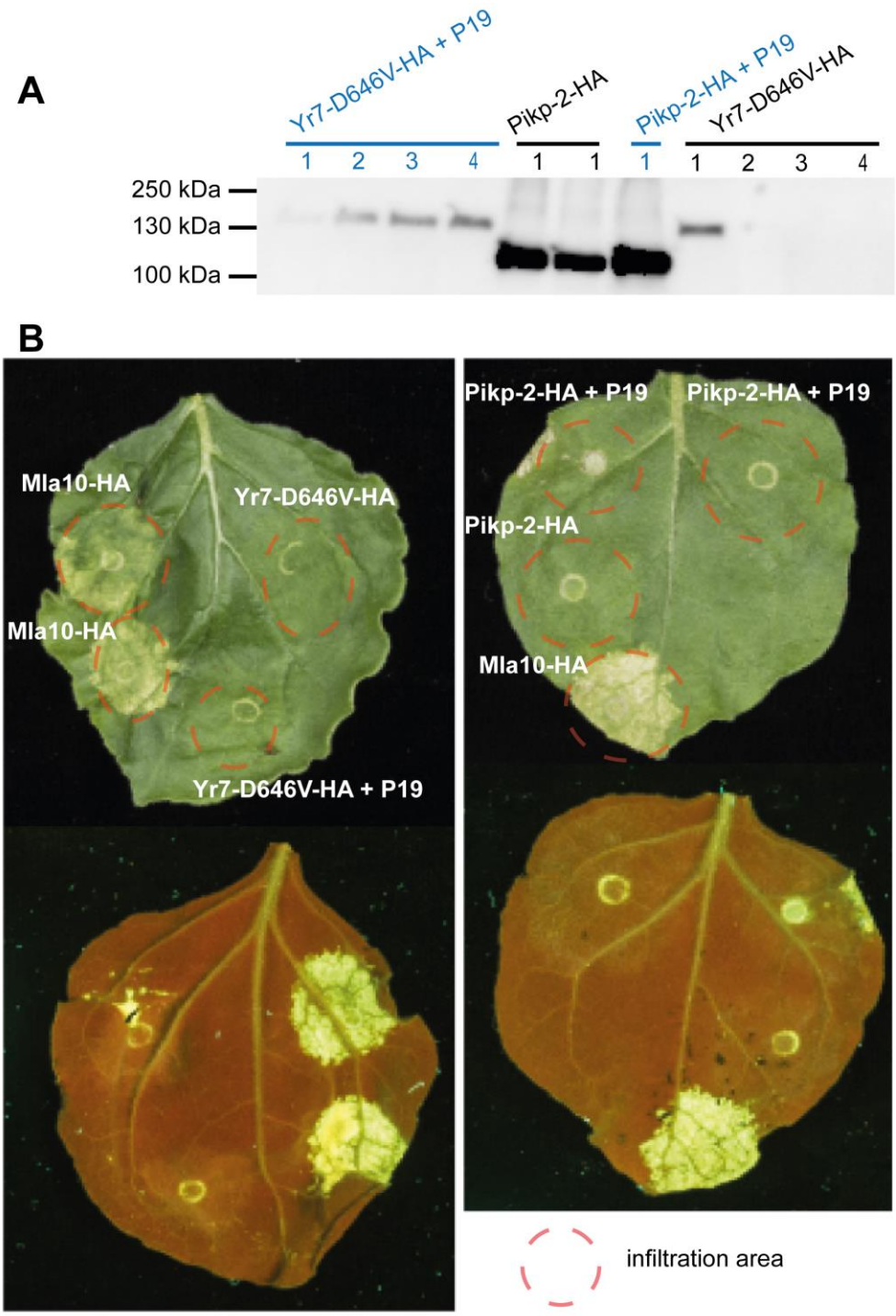


Figure 5-4. Yr7 MHD mutant expression and HR assays in *N. benthamiana*.

A: Western blots showing the protein levels of Yr7-D646V-HA co-infiltrated or not with the silencing suppressor P19 at 1, 2, 3 and 4 dpi. Pikp2-HA was used as a positive control for the Western blot.

B: HR assays (7 dpi) with Mla10 (positive control), Pikp2-HA +/- P19 (negative control) and Yr7-D646V- HA +/- P19. Top pictures were taken with normal light and bottom pictures with UV lights to emphasize the regions where cell death occurred. Note that bottom pictures are inverse to the top ones.

5.3.3.2. Truncations in Yr7

Despite the successful transient expression of *Yr7* and its mutants in the MHD motif, we did not recapitulate an HR in *N. benthamiana*. We thus have no evidence of *Yr7* being able to signal in *N. benthamiana* and this is important for further tests, including with AvYr7 candidates. We thus tested whether generating various truncations in the N-terminus of Yr7 would induce an HR when transiently expressed in *N. benthamiana*.

We generated a total of five truncations in the N-terminus of Yr7: single exon 1, single exon 2, AA201 (the first 201 residues after start codon), AA242 and AA308 (up to the NB-ARC domain). We tested their expression and capacity of inducing HR in *N. benthamiana* (Figure 5-5 and Figure 5-6). All truncations were expressed in the tested conditions, although there might be a sample overload issue for the 2 dpi samples in Figure 5-5.

None of the Yr7 truncations induced an HR in *N. benthamiana* (Figure 5-5 and Figure 5-6). However, a slightly increased signal was observed for Yr7-AA308. To validate this, we carried out additional HR assays with this truncation (Figure 5-7). Although it seemed that there was a slight increase of chlorosis in the Yr7-AA308 area on the pictures taken in normal light, it was very difficult to differentiate between the YFP signal and the Yr7-AA308-YFP signal on the UV pictures (blue arrows in Figure 5-7). It is thus unclear whether the signal observed for Yr7-AA308-YFP is relevant in these conditions.

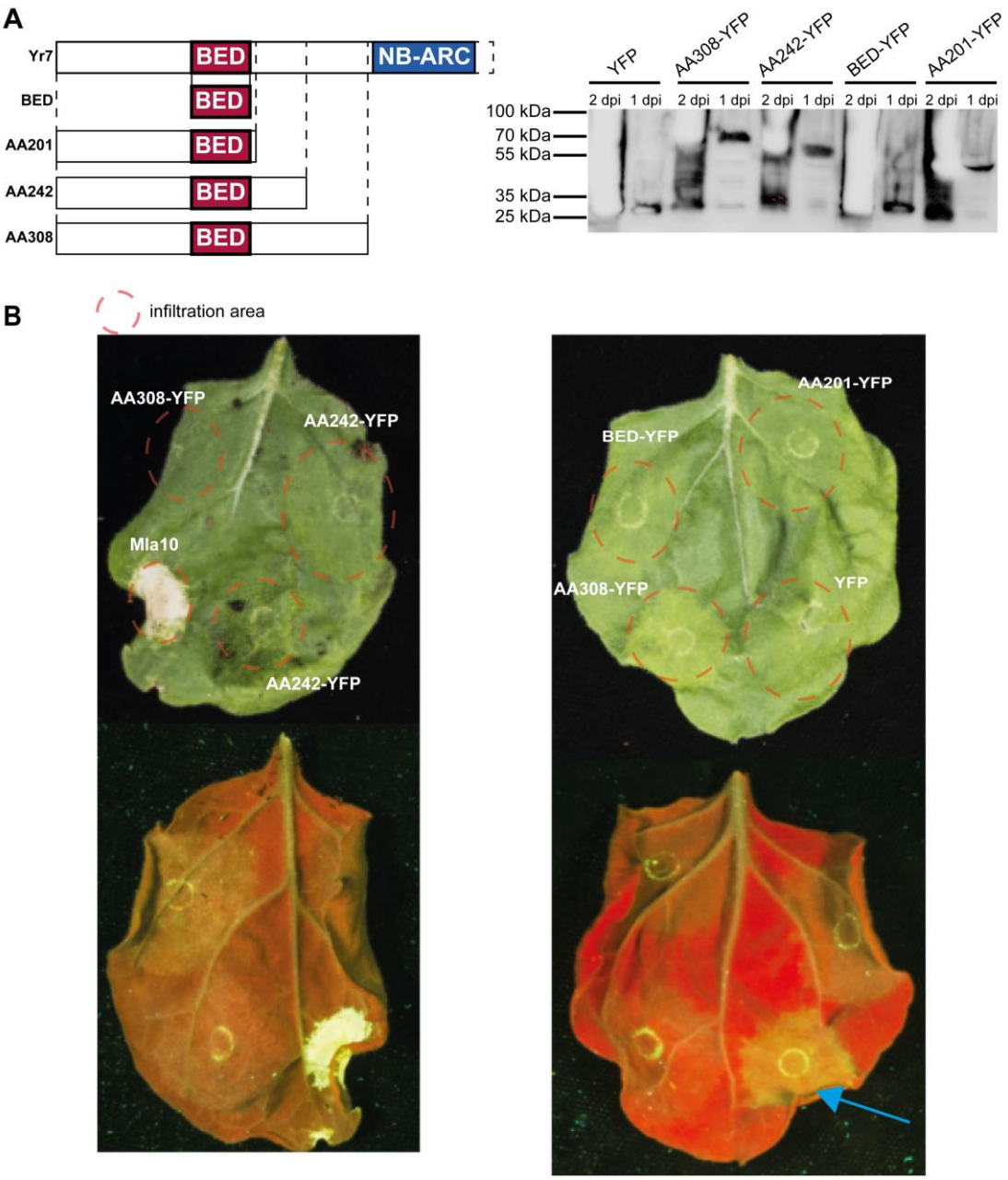


Figure 5-5. Yr7 truncations expression and HR assays in *N. benthamiana*.

A: schematics showing the boundaries of the Yr7 truncations tested on the Western Blot. Pikp2-HA was used as a positive control for the Western blot. The expression of Yr7 truncations was tested at 1 and 2 dpi, although there was a sample overload at 2 dpi.

B: HR assays with Yr7 truncations (7 dpi). Mla10 was used as a positive control and YFP as a negative control for HR. Top pictures were taken with normal light and bottom pictures with UV lights to emphasized the areas showing cell death. The blue arrow points to an area potentially showing increased signal when compared to the negative control YFP.

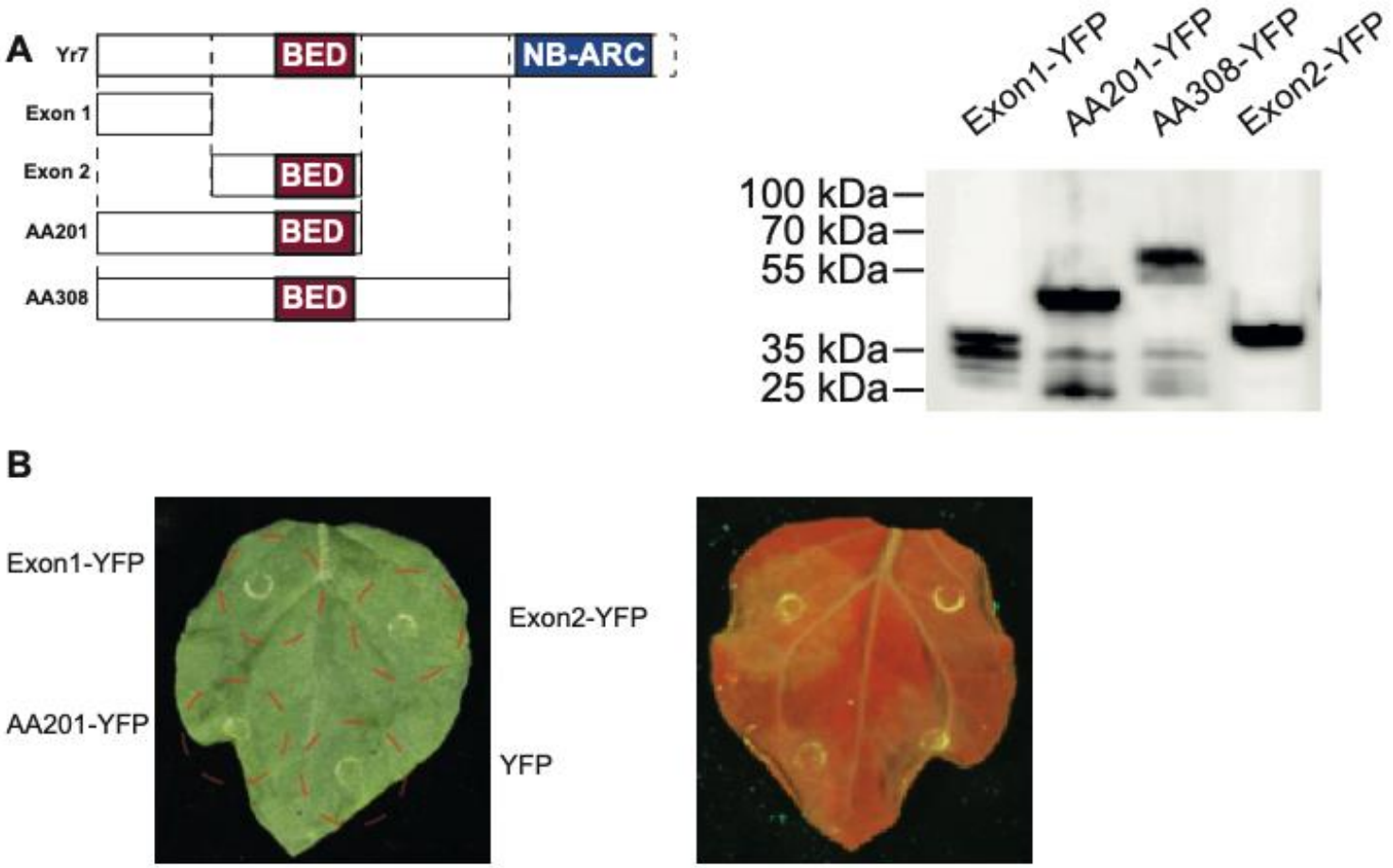


Figure 5-6. Yr7 truncations expression and HR assays in *N. benthamiana*.

A: schematics showing the boundaries of the Yr7 truncations tested on the Western Blot. The expression of Yr7 truncations was tested at 1 dpi.

B: HR assays with Yr7 truncations (7 dpi). Mla10 was used as a positive control and YFP as a negative control for HR. Top pictures were taken with normal light and bottom pictures with UV lights to emphasized the areas showing cell death.

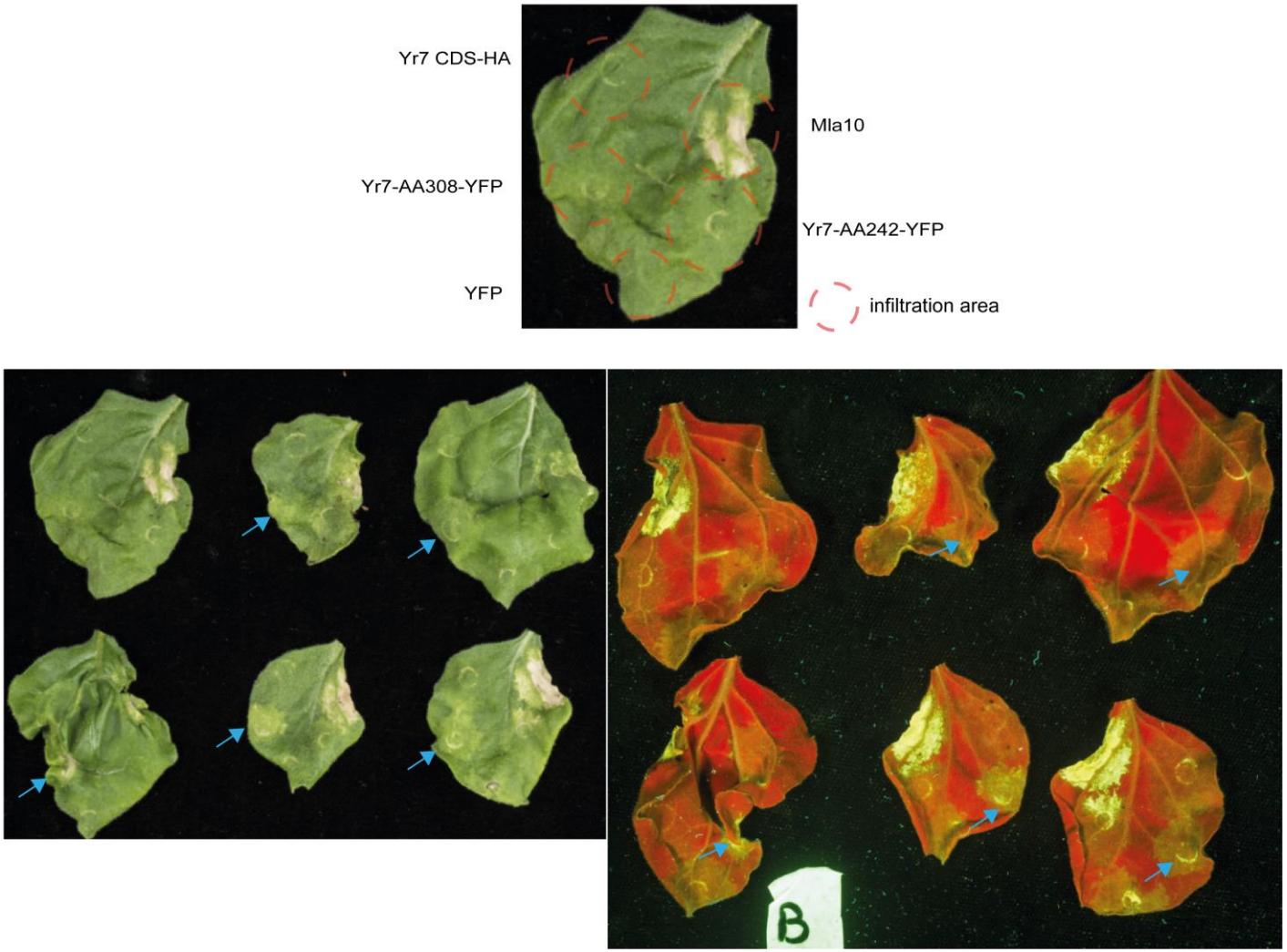


Figure 5-7. Additional HR assays with selected Yr7 truncations to confirm the slight increase in cell death signal observed in Figure 5-5 for Yr7-AA308.

Infiltrations were performed in an identical way to Figure 5-5. Pictures were taken at 7 dpi. Top pictures were taken with normal light and bottom pictures with UV lights to emphasize the regions where cell death occurred. Blue arrows point to the areas where it seems that the signal is increased in Yr7-AA308.

5.3.4. Cellular localisation of *Yr7* truncations in *N. benthamiana*

We identified a putative Nuclear Localization Signal (NLS) in Yr7 and Yr5 alleles (Chapter 4). We thus tested whether the NLS in Yr7 was functional in the Yr7 truncations (all our attempts to express Yr7 CDS-YFP failed so far). We tested Yr7-AA241 and Yr7-AA308 truncations and their corresponding deletion mutants in the NLS (Figure 5-8).

Our negative control YFP displayed a nucleo-cytoplasmic localisation, as expected. The YFP signal from both Yr7-AA242 and Yr7-AA308 truncations, which contain the predicted NLS, is located exclusively in the nucleus (Figure 5-8). The shorter recombinant protein Yr7-AA201-YFP that does not contain the NLS displayed a nucleo-cytoplasmic localisation similar to what we observed for YFP alone. The predicted NLS is thus important for the cellular localisation of Yr7-AA242 and Yr7-AA308. To confirm that the predicted is responsible for the nuclear localisation of both Yr7-AA242 and Yr7-AA308, we generated the corresponding mutants lacking the NLS (Figure 5-8). Both mutants having showed a localization similar to single YFP and AA201-YFP (Figure 5-8). Thus, NLS signal is functional in *N. benthamiana* in Yr7 truncations. Whether the full length Yr7 localises exclusively in the nucleus remains to be tested once we achieve transient expression of Yr7-CDS-YFP in *N. benthamiana*.

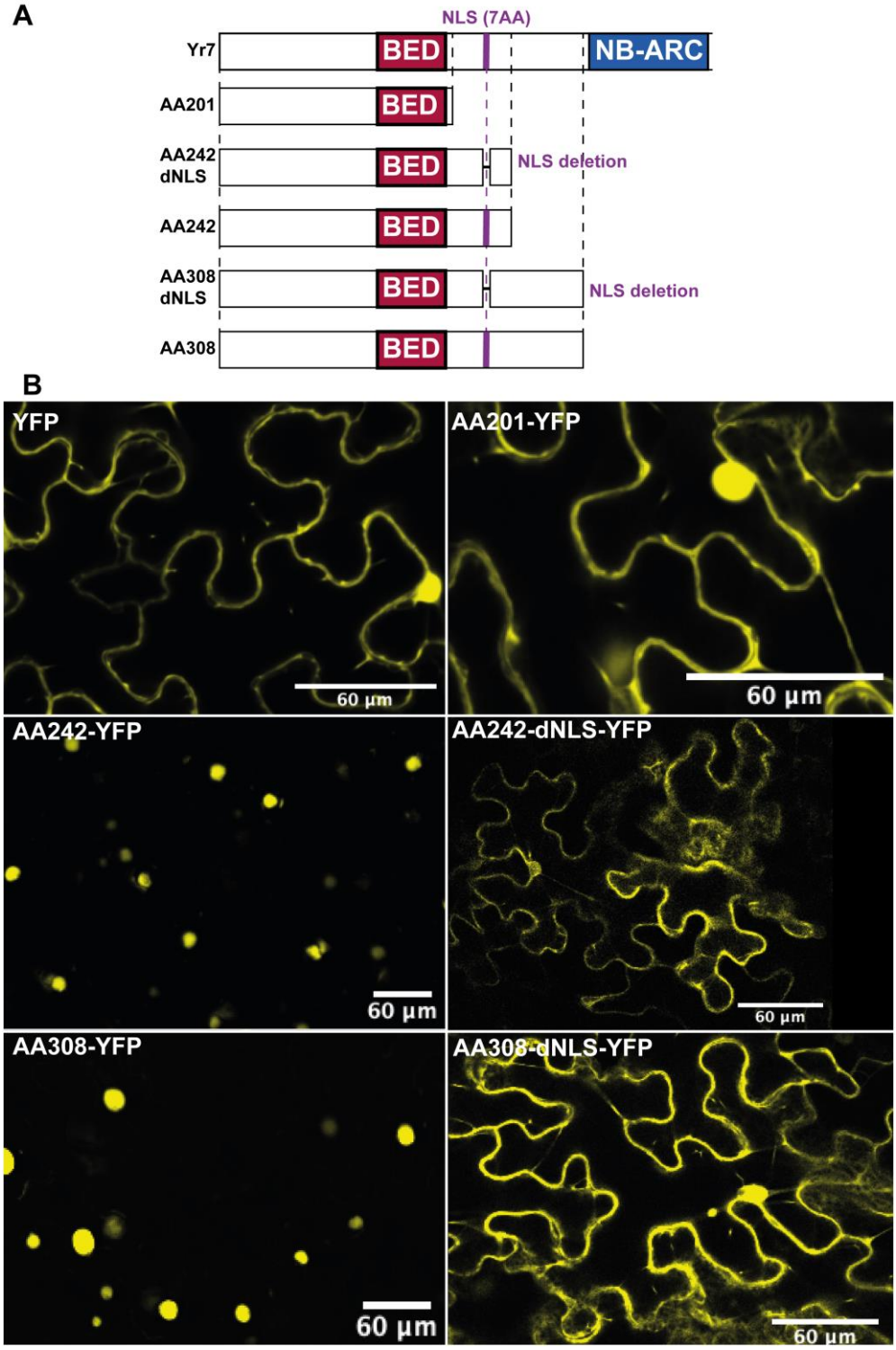


Figure 5-8. Cellular localisation of selected Yr7 truncations in *N. benthamiana* observed. Sample were taken at 1.5 dpi.
Five leaf fragments from two leaves per plant (two plants per construct) were assessed under confocal microscope (see Methods section 5.2.4). Single YFP was used as a negative control. Yr7-AA201 does not contain the predicted NLS. Both Yr7-AA242 and AA308 do contain the predicted NLS and Yr7-AA242-dNLS and Yr7-AA308-dNLS are mutants lacking the predicted NLS.

5.4. Discussion

5.4.1. Development of Fielder+Yr7 transgenic lines

We successfully generated 15 T_0 lines carrying the *Yr7* transgene and kept two lines that did not have it as negative control for further pathology tests (Table 5-1). We advanced five of these T_0 lines to T_3 seeds that are likely to be homozygous for the transgene. Transformation of *Yr5* and *YrSP* under both Sr33 and their native regulatory elements (744bp and 671bp for 5' regulatory elements and 1,500bp and 2,092bp for 3' regulatory elements for *Yr5* and *YrSP*, respectively) led to expression of the corresponding resistance in Fielder (Jianping Zhang, personal communication). We sent the material listed in Table 5-4 to our collaborator Peng Zhang (University of Sydney) to test whether *Yr7*-mediated resistance could be recapitulated in Fielder+Yr7 transgenic lines.

5.4.2. No activity detected for *Yr7* in *N. benthamiana*

In this Chapter, we aimed to test whether *N. benthamiana* was a suitable heterologous system to study *Yr7* function. We first optimised Yr7 protein expression in *N. benthamiana* and then tested different variants of the Yr7 protein with the aim to trigger HR (see section 5.1.2).

We transiently expressed *Yr7* in *N. benthamiana* (Figure 5-3), along with its variants including a MHD mutant (Figure 5-4) and N-terminus truncations (Figure 5-5 and Figure 5-6). However, we were not able to recapitulate cell-death in this system in all tested conditions, despite positive and negative controls behaving accordingly. Several reasons could explain why Yr7 or Yr7-D646V is not able to trigger HR in *N. benthamiana*:

(i) *N. benthamiana* does not possess BED-NLRs (data not shown) and may thus lack the required components involved in BED-NLR-mediated resistance. Therefore, expressing *Yr7* or its variants is not sufficient to induce HR in this system.

(ii) Although mutations in the MHD motif were shown to lead to auto-activity in *Rx*, it could be that it is not sufficient to do so in *Yr7*. Furthermore, it could be that other mutations than D to V in the MHD motif are necessary to induce auto-activity.

It is thus unclear whether *N. benthamiana* is a suitable system to study *Yr7* function with our current knowledge. Additional component involved in the *Yr7*-mediated resistance might be required to recapitulate it in *N. benthamiana*. We propose the following experiments to detect such partners:

(i) Providing the *Yr7* resistance can be recapitulated in Fielder, we plan to transform this cultivar with a recombinant Yr7 protein that is tagged to allow co-immunoprecipitation of any interactants in wheat directly and under control and disease conditions. We will couple this experiment with mass-spectrometry to identify the potential interactants and determine whether there is any difference between presence and absence of *Pst*. This will provide us with additional evidence regarding the potential mode of action of *Yr7*.

(ii) If the *Yr7* resistance cannot be recapitulated in Fielder, we plan to carry out yeast-two-hybrids assays with Yr7 against a library derived from wheat infected leaves. This would allow us to identify potential interacting partners under disease pressure. Additionally, it might also be possible to identify potential *Pst* proteins interacting with Yr7.

In a parallel experiment that is not discussed in this thesis, we identified two *AvrYr7* candidates. We are currently optimising their expression in *N. benthamiana* and these candidates are being tested via virus-induced over-expression (VOX) in Cadenza wild-type and *Yr7* loss of function mutants by Kostya Kanyuka (Rothamsted Research). Providing overexpressing one of the *AvrYr7* candidate triggers an HR in Cadenza but not in the mutants, we aim to co-express this candidate with *Yr7* in *N. benthamiana* to test for indirect interaction via co-immunoprecipitation. This would answer the question whether another partner is involved in *Yr7*-mediated resistance. Indeed, if effector recognition occurs in *N. benthamiana*, it suggests that *Yr7* is sufficient to confer resistance. If *Yr7* alone is not sufficient, then it is likely that other interacting proteins that are not present in *N. benthamiana* are required.

5.4.3. NLS identified in *Yr7* is functional in *Yr7* truncations

We identified a predicted Nuclear Localisation Signal in Yr7 (Chapter 3). We tested whether this signal was functional in *N. benthamiana* in Yr7 truncations with or without the predicted NLS (Figure 5-8). We observed that all YFP-tagged truncation harbouring the NLS were exclusively located in the nucleus (Yr7-AA242 and Yr7-AA308), whereas a smaller truncation that do not contain the predicted NLS displayed a nucleo-cytoplasmic localisation similar to YFP alone (Yr7-AA201) (Figure 5-8). Additionally, when the NLS was deleted in the Yr7-AA242 and Yr7-AA308 constructs, these mutants showed a nucleo-cytoplasmic localisation similar to what we observed for Yr7-AA201 (Figure 5-8). This suggests that the NLS is functional in *N. benthamiana* in the tested Yr7 truncations.

We could not express Yr7 full-length with a YFP tag to determine whether the full-length protein also localises exclusively in the nucleus. We are currently working on

troubleshooting this via testing other fluorescent tags. Additionally, as discussed above we aim to develop Yr7-tagged transgenic lines. We will thus develop a line with a recombinant Yr7 protein with a fluorescent tag that will allow both Co-IP and cellular localisation experiments. We also aim to test our mutants with deleted NLS in the full-length context to determine whether the nuclear localisation of Yr7 is required for the expression of resistance against *Pst*.

We discussed in Chapter 4 that nuclear localisation was required for both *Mla10* and *RPS4*-mediated resistance[269,270]. This would be consistent with our current hypothesis proposing that BED-NLRs function in a similar manner to RRS1-R. Indeed, PopP2 is able to interact with WRKY transcription factors to prevent them to bind DNA[45,231]. Additionally, it has been shown that the interaction between PopP2 and RRS1-R localise in the nucleus[313]. Providing the putative AvrYr7 targets BED-containing transcription factor(s), it is thus likely that BED-NLRs need to localise or re-localise in the nucleus during infection to be able to interact with the effector. Our planned experiments that we discussed in this section will allow us to test this hypothesis.

5.4.4. Summary

In this Chapter, we showed that we were able to transiently express *Yr7* in *N. benthamiana* and confirm the functionality of the predicted NLS in Yr7 truncations. However, we did not observe a HR either with wild-type *Yr7* or variants including MHD mutant and N-terminus truncations. It is thus unclear whether *N. benthamiana* is a suitable system to study *Yr7* function. We thus aim to carry-out CoIP-MS in Yr7-tagged wheat transgenic lines and a yeast-two-hybrids screen with *Yr7* against a cDNA library derived from wheat infected leaves to identify potential partners that interact with Yr7. We plan to use the Yr7-tagged transgenic lines to validate the functionality of the

predicted NLS and determine whether it is required for *Yr7*-mediated resistance in wheat. Additionally, we are currently testing candidates for *AvrYr7* via VOX in Cadenza wild-type and Cadenza *Yr7* loss-of-function mutants. These experiments will provide more knowledge regarding *Yr7* function in wheat directly and will allow us to determine whether *Yr7* functions in a similar manner to RRS1-R.

6. General discussion

The aim of this thesis was to understand the molecular components of yellow rust resistance in hexaploid wheat *Triticum aestivum* by focusing on specific resistance genes (*Yr7*, *Yr5*, *YrSP* and *Yr12*). To achieve this, we proposed to use forward genetics to identify *Yr7*, *Yr5*, *YrSP* and *Yr12*-loss of function mutants (Chapter 2) and carry out a MutRenSeq experiment to clone the corresponding genes (Chapter 3). We then combined comparative genomics and neighbour-net approaches to generate hypotheses regarding their mode of action (Chapter 4) and test these hypotheses (Chapter 5).

In this Chapter, we will summarise our main findings, discuss how they will provide more insight regarding BED-NLRs function in wheat and how this information could be of use in breeding programs. We will then propose future experiments to test the new hypotheses we generated regarding the mode of action of *Yr7* and *Yr5*/*YrSP*.

6.1. MutRenSeq is an effective approach to clone NLR genes in wheat

We successfully used a MutRenSeq approach to clone *Yr7*, *Yr5* and *YrSP* (Chapters 2 and 3) and identify several candidates for *Yr12*. The *Yr12* results were obtained shortly before thesis submission and were thus not included in this thesis. This method is thus a powerful tool to clone NLR genes in wheat. Indeed, it is reference-free so it overcomes the issue of mapping-to-reference-based approaches when available wheat genomes do not contain the targeted gene. Additionally, with the improvements made in the design of the bait library (to include intronic regions), the issue of fragmented contigs is largely resolved (discussed in Chapter 3). Alternatively, one might consider using long-read sequencing techniques to circumvent this[189]. Furthermore, this technique does not rely on recombination. Therefore if the gene of interest is located in a low-recombination region, which is frequent in wheat[112], it does not pose a major problem.

However, note that developing and testing the mutant population may be time-consuming, especially when seedling tests are not an option. In this thesis we showed that depending on the starting material, the associated timelines vary a lot:

- *Yr5*-loss of function mutants in the Lemhi-Yr5 background[174]: mutants were already identified and their phenotype confirmed[174]. Carrying-out the MutRenSeq analysis in this case led to the identification of a single candidate in less than six months from receiving the Lemhi-Yr5 mutant seeds to identify the candidate contig.

- *Yr7*-loss of function mutants in the Cadenza background: in this particular case, the EMS-mutagenised population in Cadenza was already available at the start of this thesis[107]. We screened two times 500 Cadenza mutant lines (1,000 lines in total) given that the first screen led to the characterisation and confirmation of only three *Yr7*-loss of function mutant lines out of the fourteen lines identified at first. We considered that this was less than what was optimal to ensure the identified candidate contigs were not obtained by chance (Chapter 1). With these two EMS screens and the following progeny tests to confirm the phenotype of the mutants, it took a longer time to identify a candidate contig for *Yr7* when compared to the *Yr5* experiment (approximately two years from the first EMS screen to the identification of the candidate).

- *Yr12*-loss of function mutants in the Armada background: This was the most 'extreme' case where we developed the EMS-mutagenised population and could not perform seedling tests to challenge it against a *Pst* isolate avirulent to *Yr12*.

We thus had to perform three field trials over three years to confirm the phenotype of the mutant lines we identified from the first field trial. We also could not predict whether Armada wild-type would remain resistant in the field over these three years and this would have severely hindered the experiment if it would not have been the case. Indeed, we could not have differentiated between Armada wild-type and the derived mutants based on yellow rust phenotype. Nevertheless, we obtained several candidate contigs for *Yr12* in November this year, four years after performing the EMS mutagenesis on Armada seeds.

Overall, our three experiments exemplify the impact of the starting material on the efficiency of the MutRenSeq approach. We will discuss in the following sections what would be an alternative way to increase the likelihood of cloning the targeted resistance gene in one single experiment with MutRenSeq.

6.1.1. Additional data on the targeted gene may increase the likelihood of obtaining only one or few candidates with MutRenSeq

We discussed in Chapter 3 that it is important to bear in mind that prior knowledge is important both before and after the MutRenSeq experiment to narrow down the list of candidates and validate the most promising ones. This knowledge includes:

- The potential nature of the gene: Given that MutRenSeq is a targeted NLR sequencing approach, strong evidence regarding the NLR nature of the gene of interest is crucial before undergoing the experiment. In our case, we had evidence showing that *Yr7*, *Yr5*, *YrSP* and *Yr12* resistance leads to a hypersensitive response, which is the hallmark of NLR-mediated resistance.

Additionally, all genes were *Pst* race-specific and dominant (Chapter 2), which is consistent with an NLR nature.

- The rough location of the gene: Depending on the initial number of independent mutant lines identified in the EMS-mutant screen, more than one gene could carry mutations in all lines (discussed in Chapter 2). Information about the chromosomal location of the targeted gene could thus be used to filter out candidates. Indeed, we showed that performing BLAST analyses in sequenced wheat genomes identified potential alleles or distant homologs for *Yr7*, *Yr5* and *YrSP*. Thus, carrying out similar analysis with all candidates could help discard the ones whose best hits are not located on the expected chromosome group.

- RNA-Seq data to validate the structure of the gene (optional): As we discussed in Chapter 3, defining the gene structure may be required depending on the follow-up studies. Briefly, if developing markers is the sole objective, then the gene structure is not needed. Indeed, we demonstrated in Chapter 3 that markers could be tested in bi-parental populations and in diversity panels and provide enough evidence to validate the candidate. However, knowing the gene structure will be mandatory if the downstream analyses require transgenic validation and further functional characterisation. Additionally, confirming that all identified mutations lead to a non-synonymous polymorphism in the predicted protein is also part of the validation process and necessitates the correct gene structure.

6.1.2. Proposed experimental design to make the most of the MutRenSeq approach

Altogether, our results suggest that one should consider the following to obtain the highest likelihood of identifying the targeted gene in one single experiment with MutRenSeq:

- Having strong evidence for the gene encoding a NLR (race-specificity, hypersensitive response, single dominance).

  This is mandatory to carry out a MutRenSeq approach.

- Developing a Near-Isogenic Line (NIL) with the targeted gene introgressed and identifying a *Pst* isolate able to discriminate between the resistant NIL and the susceptible recurrent parent.

  We understand that generating a NIL takes time because of the several rounds of backcrossing required to decrease the size of the introgressed segment as much as possible. One may thus consider using environmental conditions allowing reducing the generation time of wheat, such as 'speed breeding'[314,315]. Furthermore, providing previous mapping information is available for the gene of interest, it is likely that such material will already be available.

- Developing the EMS-mutagenised population in this Near-Isogenic Line (NIL) and challenging it with the tested *Pst* isolate. Providing a *Pst* isolate has been successfully identified to discriminate between the yellow rust disease resistance of the NIL carrying the gene of interest and the

susceptibility of the recurrent parent, challenging the EMS population with this isolate at the seedling stage and under controlled environment conditions may be an efficient way to identify corresponding loss of function mutants.

- Carrying out MutRenSeq on the mutant lines, resistant NIL parent and susceptible recurrent parent.

  This adds an extra layer of validation in addition to comparing the sequences of the mutants to the wild-type. Indeed, the targeted contig should be either absent or carry SNPs between the resistant NIL and the susceptible recurrent parent. Therefore, it can potentially help narrow down the list of candidates. For example, we obtained several candidates for *Yr12* and none of them carried mutations in all twelve submitted mutant lines. This could be explained because we do not know what other functional *Yr* genes Armada might carry apart from *Yr12*, *Yr3a* and *Yr4a*. Thus, by carrying out our pathology tests in the field directly, we could have identified loss-of-function mutants for other genes that are functional against current *Pst* races. However, because we included both a resistant and a susceptible Skyfall NIL derived from a cross with Armada that were selected with *Yr12* flanking markers, we could narrow down our list of candidates to one via comparing the candidates contig between Armada wild-type, Armada mutant, resistant Skyfall and Skyfall NILs.

We used similar materials as described above to clone both *Yr5* and *YrSP* (Lemhi-Yr5, AvocetS-Yr5 and AvocetS-YrSP NILs). Even with four mutant lines investigated in the case of AvocetS-YrSP NIL, there was only one clear candidate for *YrSP*. In the case of *Yr7*, we successfully cloned it in a single experiment using Cadenza based on preliminary

pathology tests which suggested that PST 08/21 was able to discriminate for *Yr7* in this background (Paul Fenwick, personal communication). However, we saw in Chapter 3 that obtaining a phenotypic response specific to *Yr7* in commercial cultivars using three different *Pst* isolates was not always trivial given that wheat harbours > 3,000 NLR loci[195] and we do not know what proportion confers *Pst* resistance. This is why using a NIL may be more effective in increasing the power of the analysis. Altogether our results demonstrate that MutRenSeq is an effective tool to achieve NLR gene cloning in wheat and could lead to identifying the gene(s) of interest with suitable plant materials.

6.2. Using diversity panels to validate the candidate genes

We discussed in Chapter 2 and in the section above that the likelihood that the candidate gene is in fact the targeted gene varies depending on the number of independent mutant lines carrying mutations in the candidate contig. Indeed, we determined that the probability of having a contig carrying a mutation in seven lines by chance was extremely low (P < 6.7E-8). However, this probability drops if only three lines are considered for example (P < 8.44 E-3). Further validation is thus required to ensure that the candidate is the actual gene. Such validation steps could include for example sequencing the contigs in all mutant lines to confirm the mutations and predicting their effect on the protein product. Additionally, testing for the presence of the gene of interest in different varieties known to carry when the gene structure is not known or as part of the validation process (Chapter 3). We will elaborate about this in this section.

6.2.1. Assembling panels including varieties known to carry the gene of interest

We showed in Chapter 3 that assembling a panel including varieties known to carry the gene based on both phenotypic and genotypic evidence was a suitable approach to

include in the validation process of candidate(s) (Cadenza-derivatives experiment for *Yr7* and *Yr5* and *YrSP* cultivars experiments in Chapter 3). One could thus consider using PCR and sequencing to validate the presence of the candidates in such varieties. However, this can be expensive and time-consuming. A relevant alternative would be to design gene-specific primers to target the candidate(s). To achieve this, one may use sequence information from both the RenSeq *de novo* assembly generated from wild-type reads and available wheat genome sequences. Indeed, aligning the best BLAST hits from all available wheat sequences will improve the likelihood of identifying SNPs that are specific to the candidate gene. One can thus subsequently design primers to target such polymorphisms. In our case, we designed KASP markers, as it is a relatively cost-effective and fast approach. Consequently, if the markers amplify for the positive allele in all varieties that are supposed to carry the gene of interest and for the alternate allele in all varieties that are not supposed to carry it, one both achieves validation of the candidate and gene-specific primer design for the gene of interest.

6.2.2. Investigating characterised wheat diversity panels

Several characterised diversity panels are available for wheat, its wild-relatives and progenitors (a non-comprehensive overview in available in Adamski et al., 2018[99]). This provides a valuable resource to determine the prevalence of the targeted gene in different wheat populations. Additionally, it can provide useful information about the targeted gene's performance in the field when combined with *Pst* surveillance information regarding virulence and avirulence haplotypes present in the field, (see Chapter 2 for *Yr7*). Furthermore, integrating pedigree information in a similar way to what we performed for *Yr7* and Cadenza derivatives in Chapter 3 provides additional evidence to validate the gene. Indeed, all varieties carrying the targeted *Yr7* allele displayed a parent/progeny relationship starting from the initial donor Thatcher. A more scattered

pattern for the positive *Yr7* alleles in the pedigree trees would have been suspicious in our case, as Thatcher is the only known *Yr7* source in hexaploid wheat cultivars.

In this thesis we further tested our *Yr7*, *Yr5* and *YrSP* markers in three different panels (UK AHDB recommended list and Gediflux[191] and Watkins[190] panels). The results showed that the prevalence of *Yr7* in these panels was consistent with what we know of *Yr7* breeding history and thus added more evidence confirming that the candidate gene was indeed *Yr7*.

Altogether, these results show that designing gene-specific markers targeting the candidate gene and testing them in as many wheat varieties as possible with prior knowledge about the presence/absence of the gene is a relevant approach to include in the validation steps. This is especially relevant if the gene's mode of action is unknown and additional partners may be required for the expression of resistance, which can hinder transgenic validation. Furthermore, in our opinion, such approach should be the systematically included in resistance gene cloning pipelines. Indeed, the main objective is to provide breeders with the necessary tools to select for the genes of interest in their programs and develop resistant varieties. It was thus important for us to ensure that the gene-specific marker sequences were publicly available in Marchal et al., 2018[153] and Marker Assisted Selection in Wheat website (https://maswheat.ucdavis.edu/).

6.3. Are BED domains in BED-NLRs integrated decoys?

We cloned *Yr7*, *Yr5* and *YrSP* and identified that their predicted protein products carry a non-canonical BED domain at the N-terminus (Chapter 3). We mentioned in Chapter 3 that in rice *Xa1*, conferring resistance against the bacterium *Xanthomonas oryzae* pv. *oryzae* (the causal agent of bacterial blight)[224,225], and a candidate gene for *Xo1*[226],

conferring resistance against both bacterial blight and bacterial leaf streak (latter caused by *Xanthomonas oryzae* pv. *oryzicola*) also encode BED-NLR immune receptors. We thus concluded that this particular domain organisation is effective against both fungal and bacterial pathogens. However, little is known about the mode of action of these BED-NLRs in rice. *Xa1* has been shown to be effective against a specific class of effectors, the transcription activator-like (TAL) effectors[227] (see Chapter 1) but the mechanisms of recognition itself is unknown. Zuluaga et al.,[316] proposed that TALEs target BED domains of BED-containing proteins *in planta* to enhance susceptibility and that BED domains in BED-NLRs have evolved as integrated decoy to perceive TALEs and trigger resistance. However, it is unclear whether *Pst* possess TALEs. We thus aimed to explore and test similar hypotheses regarding the mode of action of this particular NLR class in plants in the context of *Pst* resistance.

Numerous NLRs with non-canonical integrated domains have been identified in plant genomes[87,88]. Additionally, we detailed in Chapter 4 what is known about three well-characterised NLR pairs in which one of the NLR partner carries a non-canonical domain (RRS1-R/RPS4[45,231] from Arabidopsis, RGA5/RGA4[232,239] and Pik1/Pik2[234,243,244] from rice). In all three cases, the integrated domain directly interacts with the pathogen effector. This led to the postulation of the 'integrated decoy' model to explain the mode of action of these particular NLR pairs, where the integrated domain serves as a 'decoy' for the effector, mimicking the original target of the effector in the plant[211]. In Chapter 4 and 5, we thus asked the question whether the BED domains observed in Yr7, Yr5/YrSP and Xa1 were integrated decoys. To address this, we investigated the two following characteristics of the NLR pairs described above:

- The NLR pair are physically close and in head to head orientation in the genome.

- The integrated domain shares similarities with the same domain present in other proteins (e.g. PopP2 is able to bind and acetylate WRKY domains from both RRS1 and other specific WRKY containing proteins[45,231]).

6.3.1. Is there an NLR locus oriented head-to-head with *Yr7* in sequenced *Yr7* cultivars?

In the three examples above, all NLR pairs are in head-to-head orientation and in close-proximity in the genome. Taking advantage of the ten long-range assemblies available from different wheat varieties, we investigated the architecture of the *Yr* locus to determine whether a potential partner would be present in the vicinity of *Yr7* (Chapter 4). We observed that *Yr7* was part of a conserved BED-NLR cluster in *Yr7* varieties (Cadenza, Landmark, Mace, Stanley). We found an NLR locus located ~ 4 kb in the distal region of *Yr7*, but were not able to predict the full-protein product due to the presence of STOP codon in the sequence (Chapter 4). Additionally, this NLR locus was not in head-to-head orientation with *Yr7*. It is thus unlikely that it constitutes a potential partner. Given that there was expression data supporting this locus, we suggested that it could be an expressed pseudogene. We discussed in Chapter 4 that expressed pseudogenes can have a regulatory role in the transcription of their 'parent gene' (their protein-coding homolog) (reviewed in Sen and Ghosh (2013)[284] and Pink et al., (2011)[285]). However, the potential NLR pseudogene near *Yr7* only shared ~79 % identity with *Yr7* itself. It is thus unsure whether it could have a regulatory role in *Yr7* expression via these processes.

Additionally, we discussed in Chapter 4 that an NLR helper that is not necessarily located in close-proximity to *Yr7* in the genome may be required for *Yr7*-mediated resistance, similarly to what was observed in the *NRC* network of Solanaceae[287]. With our current knowledge it was not possible to test this hypothesis, although it is important to bear in mind for future experiments.

6.3.2. Do BED domains from BED-NLRs share similarities with BED domains from other BED-containing proteins?

We mentioned that PopP2 is able to bind and acetylate WRKY domains from both RRS1-R/S and other specific WRKY containing proteins and this acetylation hinders the ability of the WRKY domain to bind DNA[45,231]. Additionally, among BED-containing proteins in plant we found evidence of a BED-containing transcription factor in which knock-out mutants have strong impact on plant development[236]. This transcription factor belongs to the *daysleeper* gene family, which carries a domain similar to the hAT homodimerization domain found in hAT transposases in addition to the BED domain[236,237]. Based on this, we hypothesized that BED-NLRs could function in a similar way to RRS1-R and that BED domain from other BED-containing proteins could be virulence targets of *Pst* during infection.

To explore this, we first investigated plant proteomes containing both BED-NLRs and other BED-containing proteins. We found 18/66 explored proteomes corresponding to this criterion and all belonged to four different Angiosperm orders (Poales, Fabideae, Malpighiales and Myrtaceae), whereas 66/69 proteomes harboured BED-containing proteins but no BED-NLRs in all Viridiplantae (Chapter 5). We thus hypothesized that BED-NLRs are an Angiosperm innovation, although BED containing proteins are present in almost all sequenced Chlorophytes. Additionally, given that BED-NLRs were

identified in different Angiosperm clades, we suggested that BED-NLRs are either derived from several independent integration events or most eudicots have lost them.

Subsequently, we pursued neighbour-net analyses to determine whether BED domains from BED-NLRs shared similarities with BED domains from other BED-containing proteins. We defined five group based on phylogeny relativeness to increase the power of the study (Pooideae, Ehrhartoideae, Panicoideae, Fabideae and *Eucalyptus grandis*, Malpighiales). We generated five neighbour-networks including BED domains from BED-NLRs and BED domains from other BED-containing proteins. In each network we recorded all clades containing BED domains from BED-NLRs and investigated whether these clades also contained BED domains from other BED-containing proteins. For each BED-containing protein investigated, we retrieved the domain composition of the whole protein to determine whether a certain protein type was present more often in BED-NLRs clades than its overall representation in BED-containing protein. We identified four domains that were significantly (*p-value* < 0.01) over-represented in BED-proteins clustering with BED-NLRs compared to their proportion in BED-proteins: Dimer_Tnp_hAT (the hAT homodimerization domain discussed above), two Domain of Unknown Function (DUFs), and F-box domains. BED domains from Dimer_Tnp_hAT and F-box proteins are thus more similar to BED domains from BED-NLRs in plants in general. Furthermore, BED domains from BED-hAT proteins clustering with BED domains from BED-NLRs were found in all five investigated networks, whereas F-box proteins were found in two groups only. We thus concluded that providing the mode of action of BED-NLRs in conserved across plants and involves processes similar to what is proposes in the integrated decoy model, BED-hAT proteins have BED domains that are more similar to BED-NLRs than to any other BED-proteins. It is thus tempting to speculate that the BED domain in *Yr7* and *Yr5* could function in a similar way to WRKY

domain in RRS1-R[45,231]. Further experimental work will be required to validate or not this hypothesis and we will discuss several aspects in section 6.6.2.

6.4. *Yr7* is not active in *Nicotiana benthamiana*

We discussed in Chapter 5 that *Nicotiana benthamiana* was the most used heterologous expression system to study NLR genes with the Hypersensitive Response (HR) being the read-out of such assays. Monocot NLRs successfully signal in *N. benthamiana*, including *Mla10*[269], *Sr50*/*AvrSr50*[297] and *Sr35*/*AvrSr35*[296]. We thus set-up transient expression of *Yr7* to determine whether *Yr7* could signal in *N. benthamiana* and thus determine whether we could use this system to study *Yr7*-mediated resistance and test the hypotheses we generated above. Although we successfully expressed Yr7 and its variants including D646V mutant in the MHD motif and different truncations up the NB-ARC domain, none of these proteins produced a HR in *N. benthamiana*. It is thus unclear whether this system is suitable to study *Yr7* function.

Nevertheless, we observed an interesting phenotype when expressing different truncated versions of Yr7 with or without the NLS we identified in Chapter 4. Indeed, Yr7 truncations AA241-YFP and AA308-YFP, both carrying the NLS, were exclusively located in the nucleus, whereas AA201-YFP did not. Additionally, both mutants in AA241-YFP and AA308-YFP lacking the predicted NLS showed a nucleo-cytoplasmic location comparable to AA201-YFP and YFP alone. This experiment thus provided evidence for the NLS in Yr7 to be functional in *N. benthamiana*, although we still need to confirm the location of the full-length Yr7-YFP recombinant protein.

We identified predicted NLS in certain BED-NLRs in wheat and related grasses, as well as in Xa1[226] (Chapter 4). These NLS differed in sequence and position, but were all

located in the vicinity of the BED domain. We discussed in Chapter 4 that nuclear localisation was required for both *Mla10* and *RPS4*-mediated resistance[269,270]. This would be consistent with our current hypothesis proposing that BED-NLRs function in a similar manner to *RRS1-R*. Indeed, providing the putative *AvrYr7* targets BED-containing transcription factor, it is thus likely that BED-NLRs need to localise or re-localise in the nucleus during infection to be able to interact with the effector. We discussed several experiments to test this hypothesis in section 6.6.2.

6.5. On the value of cloning resistance genes

6.5.1. Developing gene-specific markers for Marker-Assisted Selection in breeding programs

We stated in section 6.2.2 that developing and testing gene-specific markers should be systematically performed when publishing newly cloned resistance genes. Indeed, the reason why we clone resistance genes in crops is to understand their function and allow their deployment in commercial cultivar. Both components are crucial to adapt sustainable breeding strategies.

One way to achieve this is via Marker-Assisted Selection[181,182] (MAS, described in Chapters 1 and 2), as it allows selecting plants at an early stage based on their DNA sequence without having to wait for the expression of the phenotype. It is thus important to use markers that are 100 % linked to the gene of interest to avoid selecting for false positive. Gene-specific markers thus enable overcoming this issue. However, designing such markers is not trivial. One needs to have access to as much sequence information as possible to be able to identify polymorphisms that are specific to the gene of interest. Indeed, we saw in Chapter 4 that our *Yr7* markers would select the Landmark/Mace/Stanley-Yr7 allele. It is thus important to determine whether this allele

is functional. We thus propose the following to achieve designing gene-specific markers for resistance genes in wheat, providing the gene has been validated beforehand (TILLING mutants, expression in susceptible cultivar, CRISPR-Cas9 mutants, etc):

- One should take advantage of all available wheat sequences, including the ten genomes we described in Chapter 1 and 4, to identify potential alleles or close homologs and select SNPs or other variants that are specific to the gene of interest.

- The markers should be tested in assembled panel(s) comprising varieties that are known to carry the gene and also varieties that are known to not carry it based on phenotypic and genotypic evidence.

  We recognise that it is a time-consuming task to gather information about the presence/absence of a given gene and assemble such panel. However, publicly available seed collections are available often at no cost. For example, in this thesis we retrieved seed from both the Germplasm Resource Unit at the John Innes Centre https://www.seedstor.ac.uk/ and the US National Plant Germplasm System https://npgsweb.ars-grin.gov/gringlobal/site.aspx?id=19 to assemble the *Yr7* varieties panel. Alternatively, assembling such panels for known resistance genes including varieties from breeding programs in different regions of the world could be considered. Ideally, DNA from such panels should be easily accessible to accelerate testing. This would require coordination from research programs, breeding companies and funding bodies to establish which varieties should be present and how the panel should be maintained and updated to match current needs.

- The confirmed marker(s) should be made publicly available. For example, it could be deposited in public repositories such as MASWheat (https://maswheat.ucdavis.edu/).

The number of cloned resistance genes in wheat has exponentially grown over the years with new technological improvements and related genomic resources[100]. We showed in Chapter 2 with the example of *Yr17*[152] and *Yr7* that deploying single-dominant resistance genes in the field was not sustainable. Indeed, it applies a strong selection pressure on the virulent pathogen isolates and often leads to the resistance being broken down soon after deployment. It is thus important to consider other strategies for the deployment of resistance. Such strategies include resistance gene introgression in commercial varieties, as we mentioned in Chapter 2 with the example of the *Yr5+Yr15* introgression into US cultivars as part of the UC Davis University breeding program (https://dubcovskylab.ucdavis.edu/breeding). Although this approach is the only one that is suitable for the current legislation, it is time-consuming to introgress new traits in commercial varieties. Indeed, one cross with the donor reshuffles the traits in the progeny and breeders have to re-start the selection process which includes selection for other traits. Additionally, providing the resistance gene originates from wheat wild-relative, the barrier of the species can sometimes not be overcome and crossing is thus not an option. We will discuss in the section below other approaches that could be considered when deploying resistance genes in wheat.

6.5.2. Transferring resistance gene cassettes in commercial cultivars

Combining several resistance genes in one transgenic cassette has been proposed in several reviews to address the issue of both time-consuming introgression into commercial cultivars (especially where the donor is a wild-relative) and rapid resistance

breakdown when single-dominant genes are released on their own[56,317–320]. However, this approach also comes with its limitations, including:

- *R*-gene suppression

  We mentioned in Chapter 5 the hybrid necrosis and auto-necrosis phenomenon that could occur when combining different NLR alleles in the same variety[292,293]. Furthermore, several studies reported *R*-gene suppression in wheat. For example, the rye-derived powdery mildew resistance gene *Pm8* is only expressed in wheat when a certain allele of *Pm3* is not translated[321]. Additionally, pairwise combinations of different alleles *Pm3* in F_1 hybrids and stacked transgenic wheat lines can result in suppression of *Pm3*-based resistance[291]. Ensuring that none of the genes present in the cassette would produce similar effects would thus need to be tested.

- Length of the inserted DNA

  How many resistance genes should be transferred to ensure durable resistance? One could assume that the more resistance genes are introduced, the less likely a 'super-virulence' will develop in challenged pathogen populations. However, transformation techniques are limited in the size of the insert and vector that can be transferred into a variety[322]. Wulff and Moscou[317] gave the example of a potential *R*-gene cassette comprising nine rust *R*-genes (three for each of leaf, stripe and stem rust) and two APRs that would range in size from 95 to 113 kb. One way to overcome this would be to perform sequential stacking[319]. This was showed in maize, where the authors stacked two herbicide resistance genes sequentially using a combination of zinc finger

nucleases (ZFNs) with modular 'trait landing pads' (TLPs) to direct transgene integration[323].

There have been major improvements in harnessing molecular tools for genetic engineering in crops. Although the issue of *R*-gene suppression due to the presence of certain alleles may still pose a problem, length of the DNA inserts and species/varieties that can be transformed are continuously increasing. Note that rigorous testing of the activity in all *R*-genes contained in the cassette in the recipient cultivar is crucial, as a successful transformation does not necessarily lead to a successful expression of the resistance (discussed above). Providing such tests are possible, this approach would be suitable to rapidly transform commercial varieties with *R*-genes that are critical in a given environment.

6.5.3. Understanding *R*-gene molecular mechanisms to engineer new resistances

In addition to deploying resistance genes via introgression or transformation in commercial varieties, adapting the known resistance mechanisms in breeding strategies is also a remarkable way of achieving resistance. We gave the example of how knowledge on the DNA motif targeted by TAL effectors (TALEs) enabled designing synthetic executor genes that provide resistance against multiple *Xanthomonas* strains[90–92]. Additionally, two recent studies successfully used CRISPR-Cas9-mediated genome editing to introduce mutations the cis-regulatory elements recognised by TALEs in three host sucrose transporter genes *SWEET11*, *SWEET13* and *SWEET14*, leading to resistance in an otherwise susceptible rice variety[94,95]. Remarkably, Eom et al., 2019[94] developed a portfolio of recessive resistance variants with different *SWEET* promoter sequences that are available as *R*-gene variants to use when a novel pathogen emerges[94]. This raises the

question whether achieving sustainable resistance in the field would include being one step ahead of the pathogen and developing synthetic alleles that can be rapidly deployed when needed.

In the case of NLRs with integrated domains (NLR-IDs), one approach could be identifying residues that are important for effector recognition and using this information to engineer new resistance specificities. Several studies investigated this in the *RGA4*/*RGA5* and *Pik1*/*Pik2* systems[234,298,324], including both sequence and structural information. For example, effectors *AVR1-CO39* and *AVR-PikD* from the blast fungus *M. oryzae* are sequence-unrelated but actually display a similar structure. This illustrate a limitation of our sequence-based analyses presented in Chapter 4, although structure was more difficult to analyse in a genome-wide manner.

We discussed in Chapter 3 that provided the alternative *Yr7* and *Yr5* alleles are all functional against different *Pst* isolates, this would consist of a portfolio of alleles that could be used in breeding programs. Furthermore, coupling this knowledge with the on-going development of gene editing techniques and their direct use in elite cultivars directly[223] would thus allow engineering potential new alleles and deploying them in the field in a shorter timeframe than traditional breeding. Gene editing thus has a great potential in breeding for resistance genes. Unfortunately, unless the current legislation on gene edited crops evolves, such major achievements will have a restricted impact on agricultural systems in the European Union.

6.6. Summary and future directions

6.6.1. Summary

We successfully used a MutRenSeq approach to clone *Yr7*, *Yr5* and *YrSP* and identify a candidate for *Yr12*. We proposed that *Yr5* and *YrSP* are two alleles of the same gene,

whereas *Yr7* is different gene. We developed gene-specific markers for all three loci for Marker-Assisted Selection in breeding programs. We found that *Yr7*, *Yr5* and *YrSP* encode BED-NLR immune receptors. Furthermore, combining comparative genomics across wheat cultivars and related grass species with neighbour-net analyses on BED domains from BED-NLRs and other BED-containing proteins in plants led us to hypothesize the role of integrated decoy for the BED domain in these BED-NLRs, that we illustrated in Figure 6-1. We demonstrated that the predicted NLS identified in *Yr7* was functional in *N. benthamiana,* although Yr7 itself is not active in this heterologous system in the conditions tested. We thus asked the question whether *N. benthamiana* is a suitable system to study *Yr7* function. Altogether we showed that taking advantage of new technologies and resources that are available for wheat enabled resistance gene cloning and generating/testing hypotheses regarding the function of said resistance genes.

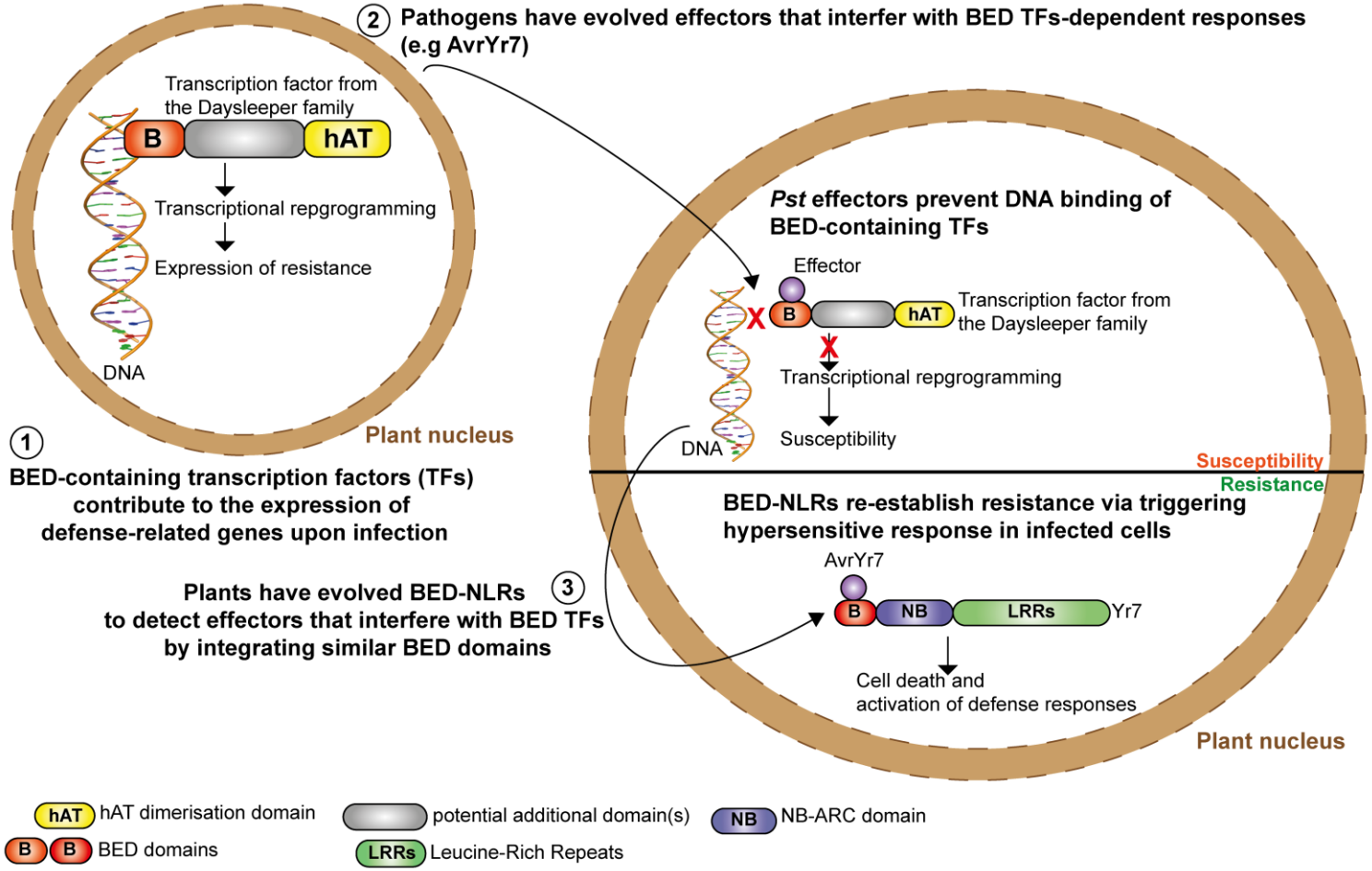


Figure 6-1. Illustration of our current hypothesis regarding the role of BED-NLRs in plant immunity.

1) BED-containing transcription factors (TFs) similar to the *daysleeper* family are involved in transcriptional reprogramming in response to pathogen that contribute to the expression of resistance.

2) Certain pathogens have evolved effectors that are able to block this response via directly or indirectly preventing DNA-binding of these TFs.

3) Certain plants have evolved BED-NLRs that are able to perceive and signal the presence of such effectors via a mechanism similar to the integrated decoy model. This re-establishes resistance in the plant.

6.6.2. Future directions

We aim to answer the following questions in the future:

Does *Yr7* require additional components to express its resistance?

We generated Fielder transgenic lines carrying *Yr7* under the regulation of *Sr33* elements. Our collaborator Peng Zhang (University of Sydney) will carry out the pathology tests to determine whether *Yr7* is functional in the Fielder background. Given that successful results were obtained with Fielder-Yr5 and -YrSP, both under the regulation of their native elements or *Sr33* elements, we expect to obtain similar outcome for Fielder-Yr7. This would suggest that either *Yr7*, *Yr5* and *YrSP* are functional on their own and do not require any additional partner, or this/these potential partner(s) are conserved across wheat varieties.

We showed in Chapter 5 that *N. benthamiana* is probably not a suitable system to study *Yr7* function with our current knowledge. Indeed, given that *N. benthamiana* does not contain BED-NLRs in its genome, it could be that it is lacking components that are important for BED-NLR signalling. We proposed the following experiments to detect such components in wheat, providing the Yr7 resistance can be recapitulated in Fielder:

(i) We will generate recombinant Yr7 protein carrying a C-terminal tag in the cultivar Fielder. This will enable co-immunoprecipitation of any interactants in wheat directly, and under control and disease conditions. We will couple this experiment with mass-spectrometry to identify the potential interactants and determine whether there is any difference between presence and absence of *Pst*. This will provide us with additional evidence regarding the potential mode of action of *Yr7*.

(ii) If the *Yr7* resistance cannot be recapitulated in Fielder, we plan to carry out yeast-two-hybrids assays with Yr7 against a library derived from wheat infected leaves. This would allow us to identify potential interacting partners under disease pressure. Additionally, it might also be possible to identify potential *Pst* proteins interacting with Yr7.

Is the NLS identified in *Yr7* important for resistance against *Pst*?

Additionally, we demonstrated that the NLS identified in *Yr7* was functional when expressing *Yr7* truncation with a YFP tag in *N. benthamiana*. To validate this localisation, we plan to generate Fielder-Yr7 lines with a fluorescent tag to enable cellular localisation experiments in wheat. Generating the same line with the NLS deleted in Yr7 will confirm whether the NLS is required for resistance against *Pst* when challenging these lines with isolate PST 08/21.

Is the BED domain an integrated decoy?

We mentioned in Chapter 5 that we conducted a parallel experiment not discussed in this thesis to identify candidates for *AvrYr7* in *Pst*. We are currently optimising expression of these candidates in *N. benthamiana*. Additionally, these candidates are being tested via virus-induced over-expression (VOX) in Cadenza wild-type and *Yr7* loss of function mutants by Kostya Kanyuka (Rothamsted Research). Obtaining HR in Cadenza, but not in the mutants, upon overexpression of one of our *AvrYr7* candidates would provide evidence for the candidate being the cognate *AvrYr7* effector. We could consequently determine which region(s) of Yr7 are important for recognition via co-expressing this candidate with *Yr7* in *N. benthamiana*. Additionally, this would answer the question whether another partner in involved in *Yr7* resistance. Indeed, if effector recognition occurs in *N. benthamiana*, it suggests that *Yr7* is sufficient to confer resistance. If *Yr7*

alone is not sufficient, then it is likely than other interacting proteins that are not present in *N. benthamiana* are required.

Further characterising the molecular function of *Yr7* in wheat will provide more insight regarding the mode of action of BED-NLRs in plants. It will also inform future hypothesis-driven engineering of novel recognition specificities to improve *Pst* resistance in the field.

7. References

1. Pardey, P. G., Beddow, J. M., Hurley, T. M., Beatty, T. K. M. & Eidman, V. R. A Bounds Analysis of World Food Futures: Global Agriculture Through to 2050. *Aust. J. Agric. Resour. Econ.* **58**, 571–589 (2014).
2. Flies, E. J., Brook, B. W., Blomqvist, L. & Buettel, J. C. Forecasting future global food demand: A systematic review and meta-analysis of model complexity. *Environ. Int.* **120**, 93–103 (2018).
3. United Nations. World Population Prospects. (2019). Available at: https://population.un.org/wpp/. (Accessed: 24th October 2019)
4. Tilman, D. *et al.* Forecasting agriculturally driven global environmental change. *Science* **292**, 281–4 (2001).
5. Godfray, H. C. J. *et al.* Food security: the challenge of feeding 9 billion people. *Science* **327**, 812–8 (2010).
6. Food and Agricultural Organization of the United Nations. State of food insecurity. (2018). Available at: http://www.fao.org/state-of-food-security-nutrition/en/. (Accessed: 24th October 2019)
7. Grassini, P., Eskridge, K. M. & Cassman, K. G. Distinguishing between yield advances and yield plateaus in historical crop production trends. *Nat. Commun.* **4**, 2918 (2013).
8. Ray, D. K., Mueller, N. D., West, P. C. & Foley, J. A. Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PLoS One* **8**, e66428 (2013).
9. Brisson, N. *et al.* Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. *F. Crop. Res.* **119**, 201–212 (2010).
10. Childe, V. G. (Vere G. *New light on the most ancient East*,. (Norton, 1969).
11. Nesbitt, M. & Samuel, D. From staple crop to extinction? The archaeology and history of the hulled wheats. in *Proceedings of the First International Workshop on Hulled Wheats: 21-22 July 1995; Castelvecchio Pascoli, Tuscany, Italy* **4**, 41–100 (1996).
12. Heun, M. *et al.* Site of Einkorn Wheat Domestication Identified by DNA Fingerprinting. *Science (80-.).* **278**, 1312–1314 (1997).
13. Dvorak, J., Luo, M.-C., Yang, Z.-L. & Zhang, H.-B. The structure of the Aegilops tauschii genepool and the evolution of hexaploid wheat. *TAG Theor. Appl. Genet.* **97**, 657–670 (1998).
14. Haider, N. The origin of the B-genome of bread wheat (Triticum aestivum L.). *Genetika* **49**, 303–14 (2013).
15. Feldman, M., Bonjean, A. & Angus, W. The Origin of cultivated wheat. *Orig. Cultiv. Wheat. Wheat B.* 1–56 (2001).
16. Kihara, H. Discovery of the DD-analyser, one of the ancestors of Triticum vulgare (abstr). *Agric. Hortic.* **19**, 889–890 (1944).
17. Dvorak, J. *et al.* The Origin of Spelt and Free-Threshing Hexaploid Wheat. *J. Hered.* **103**, 426–441 (2012).
18. Blatter, R., Jacomet, S. & Schlumbaum, A. About the origin of European spelt (Triticum spelta L.): Allelic differentiation of the HMW Glutenin B1-1 and A1-2 subunit genes. *Theor. Appl. Genet.* **108**, 360–367 (2004).
19. Nalam, V. J., Vales, M. I., Watson, C. J. W., Kianian, S. F. & Riera-Lizarazu, O. Map-based analysis of genes affecting the brittle rachis character in tetraploid wheat (Triticum turgidum L.). *Theor. Appl. Genet.* **112**, 373–381 (2006).
20. Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**, 1862–6 (2007).

21. Hay, R. K. M. Harvest index: a review of its use in plant breeding and crop physiology. *Ann. Appl. Biol.* **126**, 197–216 (1995).
22. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
23. Avni, R. *et al.* Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science (80-.).* **357**, 93–97 (2017).
24. Oerke, E. C. Crop losses to pests. *J. Agric. Sci.* **144**, 31–43 (2006).
25. Saari, E. E. & others. Distribution and importance of root rot diseases of wheat, barley and triticale in South and Southeast Asia. in *Wheats for more tropical environments. A proceedings of the international symposium, September 24-28, 1984 Mexico DF* 189–195 (1985).
26. Savary, S. *et al.* The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* 1 (2019). doi:10.1038/s41559-018-0793-y
27. Singh, R. P. *et al.* The Emergence of Ug99 Races of the Stem Rust Fungus is a Threat to World Wheat Production. *Annu. Rev. Phytopathol.* **49**, 465–481 (2011).
28. Flor, H. H. Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* **9**, 275–296 (1971).
29. Kolmer, J. A. Tracking wheat rust on a continental scale. *Curr. Opin. Plant Biol.* **8**, 441–9 (2005).
30. Milus, E. A., Kristensen, K. & Hovmøller, M. S. Evidence for increased aggressiveness in a recent widespread strain of Puccinia striiformis f. sp. tritici causing stripe rust of wheat. *Phytopathology* **99**, 89–94 (2009).
31. Wellings, C. R. Global status of stripe rust: a review of historical and current threats. *Euphytica* **179**, 129–141 (2011).
32. Beddow, J. M. *et al.* Research investment implications of shifts in the global geography of wheat stripe rust. *Nat. Plants* **1**, (2015).
33. Chen, W. Q., Wellings, C., Chen, X. M., Kang, Z. S. & Liu, T. G. Wheat stripe (yellow) rust caused by Puccinia striiformis f. sp.tritici. *Mol. Plant Pathol.* **15**, 433–446 (2014).
34. Schwessinger, B. Fundamental wheat stripe rust research in the 21 st century. *New Phytol.* (2016). doi:10.1111/nph.14159
35. Hovmøller, M. S., Sørensen, C. K., Walter, S. & Justesen, A. F. Diversity of *Puccinia striiformis* on Cereals and Grasses. *Annu. Rev. Phytopathol.* **49**, 197–217 (2011).
36. Jin, Y., Szabo, L. J. & Carson, M. Century-Old Mystery of *Puccinia striiformis* Life History Solved with the Identification of *Berberis* as an Alternate Host. *Phytopathology* **100**, 432–435 (2010).
37. Schwessinger, B. Fundamental wheat stripe rust research in the 21 st century. *New Phytol.* **213**, 1625–1631 (2017).
38. Park, R. F. & Wellings, C. R. Somatic Hybridization in the Uredinales. *Annu. Rev. Phytopathol.* **50**, 219–239 (2012).
39. Li, F. *et al.* Emergence of the Ug99 lineage of the wheat stem rust pathogen through somatic hybridisation. *bioRxiv* 692640 (2019). doi:10.1101/692640
40. Hovmøller, M. S. *et al.* Replacement of the European wheat yellow rust population by new races from the centre of diversity in the near-Himalayan region. *Plant Pathol.* **65**, 402–411 (2016).
41. Leonard, K. J. & Szabo, L. J. Stem rust of small grains and grasses caused by Puccinia graminis. *Mol. Plant Pathol.* **6**, 99–111 (2005).
42. Dangl, J. L. & Jones, J. D. G. Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826–833 (2001).
43. McHale, L., Tan, X., Koehl, P. & Michelmore, R. W. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7**, 212 (2006).

44. Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T. & Dodds, P. N. A novel conserved mechanism for plant NLR protein pairs: the 'integrated decoy' hypothesis. *Front. Plant Sci.* **5**, 10 (2014).
45. Le Roux, C. *et al.* A Receptor Pair with an Integrated Decoy Converts Pathogen Disabling of Transcription Factors to Immunity. *Cell* **161**, 1074–1088 (2015).
46. Padmanabhan, M., Cournoyer, P. & Dinesh-Kumar, S. P. The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. *Cell. Microbiol.* **11**, 191–198 (2009).
47. Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R. W. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell* **15**, 809–34 (2003).
48. Zhou, T. *et al.* Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol. Genet. Genomics* **271**, 402–15 (2004).
49. Kohler, A. *et al.* Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Mol. Biol.* **66**, 619–636 (2008).
50. Lozano, R., Hamblin, M. T., Prochnik, S. & Jannink, J.-L. Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC Genomics* **16**, 360 (2015).
51. Andolfo, G. *et al.* Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC Plant Biol.* **14**, 120 (2014).
52. Jupe, F. *et al.* 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 J.* **76**, 530–544 (2013).
53. Michelmore, R. W. & Meyers, B. C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113–30 (1998).
54. Feng, J. Y. *et al.* Molecular mapping of *YrSP* and its relationship with other genes for stripe rust resistance in wheat chromosome 2BL. *Phytopathology* **105**, 1206–1213 (2015).
55. Zhang, P., McIntosh, R. A., Hoxha, S. & Dong, C. Wheat stripe rust resistance genes *Yr5* and *Yr7* are allelic. *Theor Appl Genet* **120**, 25–29 (2009).
56. Ellis, J. G., Lagudah, E. S., Spielmeyer, W. & Dodds, P. N. The past, present and future of breeding rust resistant wheat. *Front Plant Sci* **5**, 641 (2014).
57. Singh, R. P. *et al.* Progress Towards Genetics and Breeding for Minor Genes Based Resistance to Ug99 and Other Rusts in CIMMYT High-Yielding Spring Wheat. *J. Integr. Agric.* **13**, 255–261 (2014).
58. Knott, D. R. GENES FOR STEM RUST RESISTANCE IN WHEAT VARIETIES HOPE AND H-44. *Can. J. Genet. Cytol.* **13**, 186- (1971).
59. German, S. E. & Kolmer, J. A. EFFECT OF GENE LR34 IN THE ENHANCEMENT OF RESISTANCE TO LEAF RUST OF WHEAT. *Theor. Appl. Genet.* **84**, 97–105 (1992).
60. Vanegas, C. D. G., Garvin, D. F. & Kolmer, J. A. Genetics of stem rust resistance in the spring wheat cultivar Thatcher and the enhancement of stem rust resistance by Lr34. *Euphytica* **159**, 391–401 (2008).
61. Krattinger, S. G. *et al.* A Putative ABC Transporter Confers Durable Resistance to Multiple Fungal Pathogens in Wheat. *Science (80-.).* **323**, 1360–1363 (2009).
62. Krattinger, S. G. *et al.* Abscisic acid is a substrate of the ABC transporter encoded by the durable wheat disease resistance gene *Lr34*. *New Phytol.* **223**, 853–866 (2019).
63. Fu, D. L. *et al.* A Kinase-START Gene Confers Temperature-Dependent

Resistance to Wheat Stripe Rust. *Science (80-.).* **323**, 1357–1360 (2009).
64. Gou, J.-Y. *et al.* Wheat Stripe Rust Resistance Protein WKS1 Reduces the Ability of the Thylakoid-Associated Ascorbate Peroxidase to Detoxify Reactive Oxygen Species. *Plant Cell* **27**, 1755–70 (2015).
65. Wang, S. *et al.* YR36/WKS1-mediated Phosphorylation of PsbO, an Extrinsic Member of Photosystem II, Inhibits Photosynthesis and Confers Stripe Rust Resistance in Wheat. *Mol. Plant* (2019). doi:10.1016/J.MOLP.2019.10.005
66. Lacombe, S. *et al.* Interfamily transfer of a plant pattern-recognition receptor confers broad-spectrum bacterial resistance. *Nat. Biotechnol.* **28**, 365–369 (2010).
67. Sorensen, C. K., Hovmoller, M. S., Leconte, M., Dedryver, F. & de Vallavieille-Pope, C. New Races of Puccinia striiformis Found in Europe Reveal Race Specificity of Long-Term Effective Adult Plant Resistance in Wheat. *Phytopathology* **104**, 1042–1051 (2014).
68. Jagger, L. J., Newell, C., Berry, S. T., MacCormack, R. & Boyd, L. A. The genetic characterisation of stripe rust resistance in the German wheat cultivar Alcedo. *Theor. Appl. Genet.* **122**, 723–733 (2011).
69. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).
70. Thomma, B. P. H. J., Nürnberger, T. & Joosten, M. H. A. J. Of PAMPs and effectors: the blurred PTI-ETI dichotomy. *Plant Cell* **23**, 4–15 (2011).
71. Brunner, F. *et al.* Pep-13, a plant defense-inducing pathogen-associated pattern from Phytophthora transglutaminases. *EMBO J.* **21**, 6681–6688 (2002).
72. van der Burgh, A. M. & Joosten, M. H. A. J. Plant Immunity: Thinking Outside and Inside the Box. *Trends Plant Sci.* **24**, 587–601 (2019).
73. Kourelis, J. & van der Hoorn, R. A. L. Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *Plant Cell* (2018). doi:10.1105/tpc.17.00579
74. Gómez-Gómez, L. & Boller, T. FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Mol. Cell* **5**, 1003–11 (2000).
75. Zipfel, C. Plant pattern-recognition receptors. *Trends Immunol.* **35**, 345–351 (2014).
76. Luderer, R., Takken, F. L. W., Wit, P. J. G. M. de & Joosten, M. H. A. J. Cladosporium fulvum overcomes Cf-2-mediated resistance by producing truncated AVR2 elicitor proteins. *Mol. Microbiol.* **45**, 875–884 (2002).
77. Dixon, M. S., Golstein, C., Thomas, C. M., van Der Biezen, E. A. & Jones, J. D. Genetic complexity of pathogen perception by plants: the example of Rcr3, a tomato gene required specifically by Cf-2. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 8807–14 (2000).
78. Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A. & Ellis, J. G. The Melampsora lini AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* **16**, 755–68 (2004).
79. Ravensdale, M. *et al.* Intramolecular Interaction Influences Binding of the Flax L5 and L6 Resistance Proteins to their AvrL567 Ligands. *PLoS Pathog.* **8**, e1003004 (2012).
80. Bernoux, M. *et al.* Comparative Analysis of the Flax Immune Receptors L6 and L7 Suggests an Equilibrium-Based Switch Activation Model. *Plant Cell* **28**, 146–59 (2016).
81. van der Hoorn, R. A. L. & Kamoun, S. From Guard to Decoy: a new model for perception of plant pathogen effectors. *Plant Cell* **20**, 2009–17 (2008).
82. Wang, G. *et al.* The Decoy Substrate of a Pathogen Effector and a Pseudokinase Specify Pathogen-Induced Modified-Self Recognition and Immunity in Plants.

Cell Host Microbe **18**, 285–295 (2015).
83. Lewis, J. D., Wu, R., Guttman, D. S. & Desveaux, D. Allele-Specific Virulence Attenuation of the Pseudomonas syringae HopZ1a Type III Effector via the Arabidopsis ZAR1 Resistance Protein. *PLoS Genet.* **6**, e1000894 (2010).
84. Seto, D. *et al.* Expanded type III effector recognition by the ZAR1 NLR protein using ZED1-related kinases. *Nat. Plants* **3**, 17027 (2017).
85. Wang, J. *et al.* Ligand-triggered allosteric ADP release primes a plant NLR complex. *Science* **364**, eaav5868 (2019).
86. Wang, J. *et al.* Reconstitution and structure of a plant NLR resistosome conferring immunity. *Science* **364**, eaav5870 (2019).
87. Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X. & Morel, J.-B. Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol.* **210**, 618–626 (2016).
88. Bailey, P. C. *et al.* Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biol.* **19**, 23 (2018).
89. Gu, K. *et al.* R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature* **435**, 1122–1125 (2005).
90. Hummel, A. W., Doyle, E. L. & Bogdanove, A. J. Addition of transcription activator-like effector binding sites to a pathogen strain-specific rice bacterial blight resistance gene makes it effective against additional strains and against bacterial leaf streak. *New Phytol.* **195**, 883–893 (2012).
91. Zeng, X. *et al.* Genetic engineering of the *Xa10* promoter for broad-spectrum and durable resistance to *Xanthomonas oryzae* pv. *oryzae*. *Plant Biotechnol. J.* **13**, 993–1001 (2015).
92. Römer, P., Recht, S. & Lahaye, T. A single plant resistance gene promoter engineered to recognize multiple TAL effectors from disparate pathogens. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 20526–31 (2009).
93. Boch, J., Bonas, U. & Lahaye, T. TAL effectors - pathogen strategies and plant resistance engineering. *New Phytol.* **204**, 823–832 (2014).
94. Eom, J.-S. *et al.* Diagnostic kit for rice blight resistance. *Nat. Biotechnol.* **37**, 1372–1379 (2019).
95. Oliva, R. *et al.* Broad-spectrum resistance to bacterial blight in rice using genome editing. *Nat. Biotechnol.* **37**, 1344–1350 (2019).
96. Johal, G. S. & Briggs, S. P. Reductase activity encoded by the HM1 disease resistance gene in maize. *Science* **258**, 985–7 (1992).
97. Carr, J. & Loebenstein, G. *Natural and engineered resistance to plant viruses : Part II*. (Elsevier Science & Technology, 2010).
98. Goutam, U. *et al.* Recent trends and perspectives of molecular markers against fungal diseases in wheat. *Front. Microbiol.* **6**, 14 (2015).
99. Adamski, N. M. *et al.* A roadmap for gene functional characterisation in wheat. (2018). doi:10.7287/peerj.preprints.26877v1
100. Keller, B., Wicker, T. & Krattinger, S. G. Advances in Wheat and Pathogen Genomics: Implications for Disease Control. *Annu. Rev. Phytopathol.* **56**, 67–87 (2018).
101. Uauy, C., Wulff, B. B. H. & Dubcovsky, J. Combining Traditional Mutagenesis with New High-Throughput Sequencing and Genome Editing to Reveal Hidden Variation in Polyploid Wheat. *Annu. Rev. Genet.* **51**, 435–454 (2017).
102. Borrill, P., Adamski, N. & Uauy, C. Genomics as the key to unlocking the polyploid potential of wheat. *New Phytol* **208**, 1008–1022 (2015).
103. Krasileva, K. V *et al.* Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* **14**, 19 (2013).

104. Endo, T. R. & Gill, B. S. The deletion stocks of common wheat. *J. Hered.* **87**, 295–307 (1996).
105. Mayer, K. F. X. *et al.* A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science (80-.).* **345**, 11 (2014).
106. Mascher, M. *et al.* Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.* **76**, 718–727 (2013).
107. Krasileva, K. V *et al.* Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U. S. A.* **6**, E913–E921 (2017).
108. Borrill, P., Ramirez-Gonzalez, R. & Uauy, C. expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform. *Plant Physiol.* **170**, 2172–86 (2016).
109. Clavijo, B. J. *et al.* An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* **27**, 885–896 (2017).
110. Clavijo, B. J. *et al.* W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. *bioRxiv* 110999 (2017). doi:10.1101/110999
111. Zimin, A. V., Puiu, D., Hall, R., Kingan, S. & Salzberg, S. L. The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *bioRxiv* (2017).
112. International Wheat Genome Sequencing Consortium (IWGSC), T. I. W. G. S. C. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
113. Zimin, A. V *et al.* The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *Gigascience* **6**, (2017).
114. Luo, M.-C. *et al.* Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature* **551**, 498 (2017).
115. Ling, H.-Q. *et al.* Genome sequence of the progenitor of wheat A subgenome Triticum urartu. *Nature* (2018). doi:10.1038/s41586-018-0108-0
116. Maccaferri, M. *et al.* Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat. Genet.* **51**, 885–895 (2019).
117. Ma, Z., Weining, S., Sharp, P. J. & Liu, C. Non-gridded library: a new approach for BAC (bacterial artificial chromosome) exploitation in hexaploid wheat (Triticum aestivum). *Nucleic Acids Res.* **28**, 106e – 106 (2000).
118. Uauy, C., Distelfeld, A., Fahima, T., Blechl, A. & Dubcovsky, J. A NAC Gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* **314**, 1298–301 (2006).
119. Thomson, M. J. High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breed. Biotechnol.* **2**, 195–212 (2014).
120. Allen, A. M. *et al.* Characterisation of a Wheat Breeders' Array suitable for high throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivium*). *Plant Biotechnol. J.* (2016). doi:10.1111/pbi.12635
121. Winfield, M. O. *et al.* High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* **14**, 1195–1206 (2016).
122. Wang, S. *et al.* Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796 (2014).
123. Arora, S. *et al.* Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat. Biotechnol.* **37**, 139–143 (2019).
124. Schneeberger, K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat. Rev. Genet.* **15**, 662–676 (2014).

125. Michelmore, R. W., Paran, I. & Kesseli, R. V. IDENTIFICATION OF MARKERS LINKED TO DISEASE-RESISTANCE GENES BY BULKED SEGREGANT ANALYSIS - A RAPID METHOD TO DETECT MARKERS IN SPECIFIC GENOMIC REGIONS BY USING SEGREGATING POPULATIONS. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 9828–9832 (1991).
126. Schneeberger, K. *et al.* SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6**, 550–551 (2009).
127. Austin, R. S. *et al.* Next-generation mapping of Arabidopsis genes. *Plant J.* **67**, 715–725 (2011).
128. Abe, A. *et al.* Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* **30**, 174–178 (2012).
129. Fekih, R. *et al.* MutMap plus : Genetic Mapping and Mutant Identification without Crossing in Rice. *PLoS One* **8**, 10 (2013).
130. Trick, M. *et al.* Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *BMC Plant Biol.* **12**, 14 (2012).
131. Liu, S. Z., Yeh, C. T., Tang, H. M., Nettleton, D. & Schnable, P. S. Gene Mapping via Bulked Segregant RNA-Seq (BSR-Seq). *PLoS One* **7**, 8 (2012).
132. Ramirez-Gonzalez, R. H. *et al.* RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnol J* **13**, 613–624 (2015).
133. King, R. *et al.* Mutation Scanning in Wheat by Exon Capture and Next-Generation Sequencing. *PLoS One* **10**, e0137549 (2015).
134. Sánchez-Martín, J. *et al.* Rapid gene isolation in barley and wheat by mutant chromosome sequencing. *Genome Biol.* **17**, 221 (2016).
135. Kaur Thind, A. *et al.* Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. *Nat. Publ. Gr.* (2017). doi:10.1038/nbt.3877
136. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–50 (2016).
137. Steuernagel, B. *et al.* Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nat. Biotechnol.* **34**, 652–655 (2016).
138. Sparks, C. A. & Jones*, H. D. Biolistics Transformation of Wheat. in *Methods in molecular biology (Clifton, N.J.)* **478**, 71–92 (2009).
139. Sparks, C. A., Doherty, A. & Jones, H. D. Genetic Transformation of Wheat via Agrobacterium-Mediated DNA Delivery. in *Methods in molecular biology (Clifton, N.J.)* **1099**, 235–250 (2014).
140. Hensel, G., Himmelbach, A., Chen, W., Douchkov, D. K. & Kumlehn, J. Transgene expression systems in the Triticeae cereals. *J. Plant Physiol.* **168**, 30–44 (2011).
141. Wang, Y. *et al.* Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nat. Biotechnol.* **32**, 947–951 (2014).
142. Fu, D., Uauy, C., Blechl, A. & Dubcovsky, J. RNA interference for wheat functional gene analysis. *Transgenic Res* **16**, 689–701 (2007).
143. Crossa, J. *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* **22**, 961–975 (2017).
144. Juliana, P. *et al.* Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theor. Appl. Genet.* 1–16 (2017). doi:10.1007/s00122-017-2897-1
145. Juliana, P. *et al.* Comparison of Models and Whole-Genome Profiling Approaches for Genomic-Enabled Prediction of Septoria Tritici Blotch, Stagonospora

Nodorum Blotch, and Tan Spot Resistance in Wheat. *Plant Genome* **10**, 0 (2017).
146. Klymiuk, V. *et al.* Cloning of the wheat Yr15 resistance gene sheds light on the plant tandem kinase-pseudokinase family. *Nat. Commun.* **9**, 3735 (2018).
147. Liu, W. *et al.* The Stripe Rust Resistance Gene Yr10 Encodes an Evolutionary-Conserved and Unique CC–NBS–LRR Sequence in Wheat. *Mol. Plant* **7**, 1740–1755 (2014).
148. Yuan, C. *et al.* Remapping of the stripe rust resistance gene Yr10 in common wheat. *Theor. Appl. Genet.* **131**, 1253–1262 (2018).
149. Rinaldo, A. *et al.* The Lr34 adult plant rust resistance gene provides seedling resistance in durum wheat without senescence. *Plant Biotechnol. J.* **15**, 894–905 (2017).
150. Herrera-Foessel, S. A. *et al.* Lr67/Yr46 confers adult plant resistance to stem rust and powdery mildew in wheat. *Theor. Appl. Genet.* **127**, 781–789 (2014).
151. Moore, J. W. *et al.* A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. *Nat. Genet.* **47**, 1494–8 (2015).
152. Bayles, R. A., Flath, K., Hovmøller, M. S. & de Vallavieille-Pope, C. Breakdown of the Yr17 resistance to yellow rust of wheat in northern Europe. *Agronomie* **20**, 805–811 (2000).
153. Marchal, C. *et al.* BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust. *Nature Plants* **4**, 662–668 (2018).
154. Zhang, P., McIntosh, R. A., Hoxha, S. & Dong, C. M. Wheat stripe rust resistance genes *Yr5* and *Yr7* are allelic. *Theor. Appl. Genet.* **120**, 25–29 (2009).
155. Zhan, G. *et al.* Virulence and molecular diversity of the *Puccinia striiformis* f. sp. *tritici* population in Xinjiang in relation to other regions of western China. *Plant Dis.* **100**, 99–107 (2016).
156. Hubbard, A. *et al.* Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biol.* **16**, 23 (2015).
157. Wellings, C. R. & McIntosh, R. A. *Puccinia striiformis f.sp. tritici* in Australasia: pathogenic changes during the first 10 years. *Plant Pathol.* **39**, 316–325 (1990).
158. Law, C. N. Genetic control of yellow rust resistance in *T. spelta* Album. *Plant Breed. Institute, Cambridge, Annu. Rep.* **1975**, 108–109 (1976).
159. Johnson, R. & Dyck, P. L. Resistance to yellow rust in *Triticum spelta* var. Album and bread wheat cultivars Thatcher and Lee. *Colloq. l'INRA* (1984).
160. McIntosh, R. A., Wellings, C. R. & Park, R. F. *Wheat rusts: an atlas of resistance genes*. (Csiro Publishing, 1995).
161. Wan, A. *et al.* Wheat Stripe Rust Epidemic and Virulence of *Puccinia striiformis* f. sp. *tritici* in China in 2002. *Plant Dis.* **88**, 896–904 (2004).
162. Ochoa, J. B., Danial, D. L. & Paucar, B. Virulence of wheat yellow rust races and resistance genes of wheat cultivars in Ecuador. *Euphytica* **153**, 287–293 (2007).
163. Johnson, R., Stubbs, R. W., Fuchs, E. & Chamberlain, N. H. Nomenclature for physiologic races of Puccinia striiformis infecting wheat. *Trans. Br. Mycol. Soc.* **58**, 475–480 (1972).
164. Schneeberger, K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat. Rev. Genet.* **15**, 662–676 (2014).
165. Tsai, H. *et al.* Production of a high-efficiency TILLING population through polyploidization. *Plant Physiol.* **161**, 1604–14 (2013).
166. Wang, W. *et al.* Gene editing and mutagenesis reveal inter-cultivar differences and additivity in the contribution of TaGW2 homoeologues to grain size and weight in wheat. *Theor. Appl. Genet.* **131**, 2463–2475 (2018).
167. Uauy, C. *et al.* A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. *BMC Plant Biol* **9**, 115 (2009).
168. Büschges, R. *et al.* The Barley Mlo Gene: A Novel Control Element of Plant

Pathogen Resistance. *Cell* **88**, 695–705 (1997).
169. Acevedo-Garcia, J. *et al. mlo* -based powdery mildew resistance in hexaploid bread wheat generated by a non-transgenic TILLING approach. *Plant Biotechnol. J.* **15**, 367–378 (2017).
170. Feuillet, C. *et al.* Map-based isolation of the leaf rust disease resistance gene Lr10 from the hexaploid wheat (Triticum aestivum L.) genome. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15253 (2003).
171. Faris, J. D. *et al.* A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. *Proc. Natl. Acad. Sci.* **107**, 13544–13549 (2010).
172. Periyannan, S. *et al.* The gene *Sr33*, an ortholog of barley *Mla* genes, encodes resistance to wheat stem rust race Ug99. *Science (80-.).* **341**, 786–788 (2013).
173. Hubbard, A. J., Fanstone, V. & Bayles, R. A. *UKCPVS 2009 Annual report*.
174. McGrann, G. R. D. *et al.* Genomic and genetic analysis of the wheat race-specific yellow rust resistance gene *Yr5*. *J. Plant Sci. Mol. Breed.* **3**, (2014).
175. Gassner, G. & Straib, W. *Die Bestimmung der biologischen Rassen des Weizengelbrostes* Puccinia glumarum f.sp. tritici *Schmidt Erikss. u. Henn*. (1932).
176. Marchal, C. *et al.* BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust. *Nat. Plants* **4**, 662–668 (2018).
177. Steuernagel, B., Witek, K., Jones, J. D. G. & Wulff, B. B. H. MutRenSeq: A Method for Rapid Cloning of Plant Disease Resistance Genes. in 215–229 (Humana Press, New York, NY, 2017). doi:10.1007/978-1-4939-7249-4_19
178. Wang, T. L., Uauy, C., Robson, F. & Till, B. TILLING *in extremis*. *Plant Biotechnol. J.* **10**, 761–772 (2012).
179. Rakszegi, M. *et al.* Diversity of agronomic and morphological traits in a mutant population of bread wheat studied in the Healthgrain program. *Euphytica* **174**, 409–421 (2010).
180. Feng, J. Y. *et al.* Molecular mapping of *YrSP* and its relationship with other genes for stripe rust resistance in wheat chromosome 2BL. *Phytopathology* **105**, 1206–1213 (2015).
181. Collard, B. C. . & Mackill, D. J. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 557–572 (2008).
182. Cobb, J. N., Biswas, P. S. & Platten, J. D. Back to the future: revisiting MAS as a tool for modern plant breeding. *Theor. Appl. Genet.* **132**, 647–667 (2019).
183. Allen, A. M. *et al.* Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.* **15**, 390–401 (2017).
184. Cavanagh, C. R. *et al.* Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8057–62 (2013).
185. Wang, L. *et al.* Large-scale identification and functional analysis of NLR genes in blast resistance in the Tetep rice genome sequence. *Proc. Natl. Acad. Sci. U. S. A.* 201910229 (2019). doi:10.1073/pnas.1910229116
186. Michelmore, R. W. & Meyers, B. C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113–30 (1998).
187. Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J. & Anderson, L. K. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.* **8**, 77–84 (2007).
188. Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G. & Wulff, B. B. H. NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics* **31**, 1665–1667

(2015).
189. Witek, K. *et al.* Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. *Nat. Biotechnol.* **34**, 656–660 (2016).
190. Wingen, L. U. *et al.* Establishing the A. E. Watkins landrace cultivar collection as a resource for systematic gene discovery in bread wheat. *Theor. Appl. Genet.* **127**, 1831–1842 (2014).
191. Reeves, J. C. *et al.* Changes over time in the genetic diversity of four major European crops - a report from the Gediflux Framework 5 project. *Genet. Var. plant breeding. Proc. 17th EUCARPIA Gen. Congr. Tulln, Austria, 8-11 Sept. 2004* 3–7 (2004).
192. Lagudah, E. S., Appels, R., Brown, A. H. D. & McNeil, D. The molecular–genetic analysis of *Triticum tauschii*, the D-genome donor to hexaploid wheat. *Genome* **34**, 375–386 (1991).
193. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
194. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677 (2013).
195. Steuernagel, B. *et al.* Physical and transcriptional organisation of the bread wheat intracellular immune receptor repertoire. *bioRxiv* 339424 (2018). doi:10.1101/339424
196. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
197. Dobon, A., Bunting, D. C. E., Cabrera-Quio, L. E., Uauy, C. & Saunders, D. G. O. The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression. *BMC Genomics* **17**, 380 (2016).
198. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
199. Lupas, A., Dyke, M. Van & Stock, J. Predicting coiled coils from protein sequences. *Science (80-.).* **252**, 1162–1164 (1991).
200. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
201. Pallotta, M. A. *et al.* Marker assisted wheat breeding in the southern region of Australia. in *Proceedings of 10th International Wheat Genet Symposium Instituto Sperimentale per la Cerealcoltura Rome* 789–791 (2003).
202. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).
203. Sun, Q., Wei, Y., Ni, Z., Xie, C. & Yang, T. Microsatellite marker for yellow rust resistance gene *Yr5* in wheat introgressed from spelt wheat. *Plant Breed.* **121**, 539–541 (2002).
204. Yao, Z. J. *et al.* The molecular tagging of the yellow rust resistance gene *Yr7* in wheat transferred from differential host Lee using microsatellite markers. *Sci. Agric. Sin.* **39**, 1146–1152 (2006).
205. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
206. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).
207. Pearce, S. *et al.* Regulation of Zn and Fe transporters by the *GPC1* gene during early wheat monocarpic senescence. *BMC Plant Biol.* **14**, 368 (2014).
208. Shaw, P. D., Graham, M., Kennedy, J., Milne, I. & Marshall, D. F. Helium: visualization of large scale plant pedigrees. *BMC Bioinformatics* **15**, 259 (2014).

209. Brunner, S. *et al.* Intragenic allele pyramiding combines different specificities of wheat *Pm3* resistance alleles. *Plant J.* **64**, 433–445 (2010).
210. Ellis, J. G., Lawrence, G. J., Luck, J. E. & Dodds, P. N. Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* **11**, 495–506 (1999).
211. Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T. & Dodds, P. N. A novel conserved mechanism for plant NLR protein pairs: the integrated decoy hypothesis. *Front. Plant Sci.* **5**, 606 (2014).
212. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
213. Bai, S. *et al.* Structure-function analysis of barley NLR immune receptor MLA10 reveals its cell compartment specific activity in cell death and disease resistance. *PLoS Pathog.* **8**, e1002752 (2012).
214. Srichumpa, P., Brunner, S., Keller, B. & Yahiaoui, N. Allelic series of four powdery mildew resistance genes at the Pm3 locus in hexaploid bread wheat. *Plant Physiol.* **139**, 885–895 (2005).
215. Bai, S. *et al.* Structure-function analysis of barley NLR immune receptor MLA10 reveals its cell compartment specific activity in cell death and disease resistance. *PLoS Pathog.* **8**, e1002752 (2012).
216. Altschul, S. F. *et al.* Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **272**, 5101–9 (2005).
217. Warren, R. F., Henk, A., Mowery, P., Holub, E. & Innes, R. W. A mutation within the leucine-rich repeat domain of the arabidopsis disease resistance gene RPS5 partially suppresses multiple bacterial and downy mildew resistance genes. *Plant Cell* **10**, 1439–1452 (1998).
218. Hofmann, K. & Stoffel, W. TMbase - A database of membrane spanning proteins segments. (1993).
219. McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallogr. D. Biol. Crystallogr.* **67**, 386–94 (2011).
220. McIntosh, R. A., Luig, N. H., Johnson, R. & Hare, R. A. Cytogenetical studies in wheat XI. Sr9g for reaction to *Puccinia graminis tritici* . *Zeitschrift fur Pflanzenzuchtung* **87**, 274–289 (1981).
221. Ramírez-González, R. H. *et al.* The transcriptional landscape of polyploid wheat. *Science (80-.).* **361**, eaar6089 (2018).
222. Sørensen, C. K., Thach, T. & Hovmøller, M. S. Evaluation of Spray and Point Inoculation Methods for the Phenotyping of *Puccinia striiformis* on Wheat. *Plant Dis.* **100**, 1064–1070 (2016).
223. Kelliher, T. *et al.* One-step genome editing of elite crop germplasm during haploid induction. *Nat. Biotechnol.* **37**, 287–292 (2019).
224. Yoshimura, S. *et al.* Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1663–8 (1998).
225. Das, B., Sengupta, S., Prasad, M. & Ghose, T. K. Genetic diversity of the conserved motifs of six bacterial leaf blight resistance genes in a set of rice landraces. *BMC Genet.* **15**, 82 (2014).
226. Read, A. C. *et al.* Genome assembly and characterization of a complex zfBED-NLR gene-containing disease resistance locus in Carolina Gold Select rice with Nanopore sequencing. *bioRxiv* 675678 (2019). doi:10.1101/675678
227. Ji, Z. *et al.* Interfering TAL effectors of Xanthomonas oryzae neutralize R-gene-mediated plant disease resistance. *Nat. Commun.* **7**, 13435 (2016).

228. Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D. G. & Krasileva, K. V. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol.* **14**, 8 (2016).
229. Bailey, P. C. *et al.* Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *bioRxiv* 100834 (2017). doi:10.1101/100834
230. Baggs, E., Dagdas, G. & Krasileva, K. NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Curr. Opin. Plant Biol.* **38**, 59–67 (2017).
231. Sarris, P. F. *et al.* A plant immune receptor detects pathogen effectors that target WRKY transcription factors. *Cell* **161**, 1089–1100 (2015).
232. Cesari, S. *et al.* The rice resistance protein pair RGA4/RGA5 recognizes the *Magnaporthe oryzae* effectors AVR-Pia and AVR1-CO39 by direct binding. *Plant Cell* **25**, 1463–1481 (2013).
233. Maqbool, A. *et al.* Structural basis of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor. *Elife* **4**, (2015).
234. De la Concepcion, J. C. *et al.* Polymorphic residues in rice NLRs expand binding and response to effectors of the blast pathogen. *Nat. Plants* 1 (2018). doi:10.1038/s41477-018-0194-x
235. Aravind, L. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem. Sci.* **25**, 421–423 (2000).
236. Bundock, P. & Hooykaas, P. An *Arabidopsis* hAT-like transposase is essential for plant development. *Nature* **436**, 282–284 (2005).
237. Knip, M., de Pater, S. & Hooykaas, P. J. The *SLEEPER* genes: a transposase-derived angiosperm-specific gene family. *BMC Plant Biol.* **12**, 192 (2012).
238. Sinzelle, L., Izsvák, Z. & Ivics, Z. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cell. Mol. Life Sci.* **66**, 1073–1093 (2009).
239. Cesari, S. *et al.* The NB-LRR proteins RGA4 and RGA5 interact functionally and physically to confer disease resistance. *Embo J.* **33**, 1941–1959 (2014).
240. Williams, S. J. *et al.* Structural Basis for Assembly and Function of a Heterodimeric Plant Immune Receptor. *Science (80-.).* **344**, (2014).
241. Wu, C.-H., Krasileva, K. V., Banfield, M. J., Terauchi, R. & Kamoun, S. The "sensor domains" of plant NLR proteins: more than decoys? *Front. Plant Sci.* **6**, 134 (2015).
242. Okuyama, Y. *et al.* A multifaceted genomics approach allows the isolation of the rice Pia-blast resistance gene consisting of two adjacent NBS-LRR protein genes. *Plant J.* **66**, 467–479 (2011).
243. Ashikawa, I. *et al.* Two Adjacent Nucleotide-Binding Site–Leucine-Rich Repeat Class Genes Are Required to Confer Pikm-Specific Rice Blast Resistance. *Genetics* **180**, 2267–2276 (2008).
244. Yoshida, K. *et al.* Association Genetics Reveals Three Novel Avirulence Genes from the Rice Blast Fungal Pathogen Magnaporthe oryzae. *Plant Cell Online* **21**, (2009).
245. Gassmann, W., Hinsch, M. E. & Staskawicz, B. J. The Arabidopsis RPS4 bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. *Plant J.* **20**, 265–277 (1999).
246. Narusaka, M. *et al. RRS1* and *RPS4* provide a dual *Resistance*- gene system against fungal and bacterial pathogens. *Plant J.* **60**, 218–226 (2009).
247. Asai, T. *et al.* MAP kinase signalling cascade in Arabidopsis innate immunity. *Nature* **415**, 977–983 (2002).

248. Fujisaki, K. *et al.* An unconventional NOI/RIN4 domain of a rice NLR protein binds host EXO70 protein to confer fungal immunity. *bioRxiv* (2017). doi:10.1101/239400
249. Aravind, L. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem. Sci.* **25**, 421–423 (2000).
250. Knip, M. *et al.* DAYSLEEPER: a nuclear and vesicular-localized protein that is expressed in proliferating tissues.
251. Takagi, H. *et al.* Rice blast resistance gene Pii is controlled by a pair of NBS-LRR genes Pii-1 and Pii-2. *bioRxiv* 227132 (2017). doi:10.1101/227132
252. Bush, S. J. *et al.* Presence–Absence Variation in A. thaliana Is Primarily Associated with Genomic Signatures Consistent with Relaxed Selective Constraints. *Mol. Biol. Evol.* **31**, 59–69 (2014).
253. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).
254. Zhu, T. *et al.* Improved Genome Sequence of Wild Emmer Wheat Zavitan with the Aid of Optical Maps. *G3 (Bethesda).* g3.200902.2018 (2019). doi:10.1534/g3.118.200902
255. Luo, M.-C. *et al.* Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498 (2017).
256. Wickham, H. *Ggplot2 : elegant graphics for data analysis*. (Springer, 2009).
257. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
258. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
259. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
260. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
261. Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive yool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067 (2012).
262. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
263. Kriventseva, E. V *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
264. Bryant, D. & Moulton, V. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265 (2003).
265. Kloepper, T. H. & Huson, D. H. Drawing explicit phylogenetic networks and their integration into SplitsTree. *BMC Evol. Biol.* **8**, 22 (2008).
266. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
267. Bernhofer, M. *et al.* NLSdb-major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Res.* **46**, D503–D508 (2018).
268. Ellis, J. G. Integrated decoys and effector traps: how to catch a plant pathogen. *BMC Biol.* **14**, 13 (2016).
269. Shen, Q.-H. *et al.* Nuclear Activity of MLA Immune Receptors Links Isolate-Specific and Basal Disease-Resistance Responses. *Science (80-.).* **315**, 1098–1103 (2007).

270. Wirthmueller, L., Zhang, Y., Jones, J. D. G. & Parker, J. E. Nuclear Accumulation of the Arabidopsis Immune Receptor RPS4 Is Necessary for Triggering EDS1-Dependent Defense. *Curr. Biol.* **17**, 2023–2029 (2007).
271. Chang, C. *et al.* Barley MLA Immune Receptors Directly Interfere with Antagonistically Acting Transcription Factors to Initiate Disease Resistance Signaling. *Plant Cell* **25**, 1158–1173 (2013).
272. Eric B. Holub. THE ARMS RACE IS ANCIENT HISTORY IN ARABIDOPSIS , THE WILDFLOWER. **516**, (2001).
273. Barakat, A., Matassi, G. & Bernardi, G. Distribution of genes in the genome of Arabidopsis thaliana and its implications for the genome organization of plants. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 10044–9 (1998).
274. Rustenholz, C. *et al.* Specific patterns of gene space organisation revealed in wheat by using the combination of barley and wheat genomic resources. *BMC Genomics* **11**, 714 (2010).
275. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
276. Schnable, P. S. *et al.* The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science (80-.).* **326**, 1112–1115 (2009).
277. Seo, E., Kim, S., Yeom, S.-I. & Choi, D. Genome-wide Comparative Analyses Reveal the Dynamic Evolution of Nucleotide-Binding Leucine-Rich Repeat Gene Family among Solanaceae Plants. *Front. Plant Sci.* **7**, 1205 (2016).
278. Zhao, Y. *et al.* Bioinformatics Analysis of NBS-LRR Encoding Resistance Genes in Setaria italica. *Biochem. Genet.* **54**, 232–248 (2016).
279. Wei, H., Li, W., Sun, X., Zhu, S. & Zhu, J. Systematic Analysis and Comparison of Nucleotide-Binding Site Disease Resistance Genes in a Diploid Cotton Gossypium raimondii. *PLoS One* **8**, e68435 (2013).
280. Kuang, H., Woo, S.-S., Meyers, B. C., Nevo, E. & Michelmore, R. W. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell* **16**, 2870–2894 (2004).
281. Chavan, S., Gray, J. & Smith, S. M. Diversity and evolution of Rp1 rust resistance genes in four maize lines. *Theor. Appl. Genet.* **128**, 985–998 (2015).
282. Prade, V. M. *et al.* The pseudogenes of barley. *Plant J.* **93**, 502–514 (2018).
283. Pingault, L. *et al.* Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biol.* **16**, 29 (2015).
284. Sen, K. & Ghosh, T. C. Pseudogenes and their composers: delving in the 'debris' of human genome. *Brief. Funct. Genomics* **12**, 536–547 (2013).
285. Pink, R. C. *et al.* Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* **17**, 792–798 (2011).
286. Wu, C.-H., Derevnina, L. & Kamoun, S. Receptor networks underpin plant immunity. *Science* **360**, 1300–1301 (2018).
287. Wu, C.-H. *et al.* NLR network mediates immunity to diverse plant pathogens. *Proc. Natl. Acad. Sci. U. S. A.* 201702041 (2017). doi:10.1073/pnas.1702041114
288. Thind, A. K. *et al.* Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars. *Genome Biol.* **19**, 104 (2018).
289. Saintenac, C. *et al.* Identification of wheat gene Sr35 that confers resistance to Ug99 stem rust race group. *Science* **341**, 783–786 (2013).
290. Mago, R. *et al.* The wheat Sr50 gene reveals rich diversity at a cereal disease resistance locus. *Nat. plants* **1**, 15186 (2015).
291. Stirnweis, D. *et al.* Suppression among alleles encoding nucleotide-binding-leucine-rich repeat resistance proteins interferes with resistance in F_1 hybrid and

allele-pyramided wheat plants. *Plant J.* **79**, 893–903 (2014).
292. Bomblies, K. *et al.* Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *Plos Biol.* **5**, 1962–1972 (2007).
293. Krüger, J. *et al.* A tomato cysteine protease required for Cf-2-dependent disease resistance and suppression of autonecrosis. *Science* **296**, 744–7 (2002).
294. Goodin, M. M., Zaitlin, D., Naidu, R. A. & Lommel, S. A. *Nicotiana benthamiana* : Its History and Future as a Model for Plant–Pathogen Interactions. *Mol. Plant-Microbe Interact.* **21**, 1015–1026 (2008).
295. Casey, L. W. *et al.* The CC domain structure from the wheat stem rust resistance protein Sr33 challenges paradigms for dimerization in plant NLR proteins. *Proc. Natl. Acad. Sci.* 201609922 (2016). doi:10.1073/pnas.1609922113
296. Salcedo, A. *et al.* Variation in the AvrSr35 gene determines Sr35 resistance against wheat stem rust race Ug99. *Science* **358**, 1604–1606 (2017).
297. Chen, J. *et al.* Loss of AvrSr50 by somatic exchange in stem rust leads to virulence for Sr50 resistance in wheat. *Science* **358**, 1607–1610 (2017).
298. Maqbool, A. *et al.* Structural basis of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor. *Elife* **4**, (2015).
299. Adachi, H., Derevnina, L. & Kamoun, S. NLR singletons, pairs, and networks: evolution, assembly, and regulation of the intracellular immunoreceptor circuitry of plants. *Curr. Opin. Plant Biol.* **50**, 121–131 (2019).
300. Boyes, D. C., Nam, J., Dangl, J. L. & Innes, R. W. The Arabidopsis thaliana RPM1 disease resistance gene product is a peripheral plasma membrane protein that is degraded coincident with the hypersensitive response. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 15849–54 (1998).
301. Qi, D., DeYoung, B. J. & Innes, R. W. Structure-function analysis of the coiled-coil and leucine-rich repeat domains of the RPS5 disease resistance protein. *Plant Physiol.* **158**, 1819–32 (2012).
302. Baudin, M., Hassan, J. A., Schreiber, K. J. & Lewis, J. D. Analysis of the ZAR1 Immune Complex Reveals Determinants for Immunity and Molecular Interactions. *Plant Physiol.* **174**, 2038–2053 (2017).
303. van Ooijen, G. *et al.* Structure–function analysis of the NB-ARC domain of plant disease resistance proteins. *J. Exp. Bot.* **59**, 1383–1397 (2008).
304. Bendahmane, A., Farnham, G., Moffett, P. & Baulcombe, D. C. Constitutive gain-of-function mutants in a nucleotide binding site-leucine rich repeat protein encoded at the Rx locus of potato. *Plant J.* **32**, 195–204 (2002).
305. Williams, S. J. *et al.* An Autoactive Mutant of the M Flax Rust Resistance Protein Has a Preference for Binding ATP, Whereas Wild-Type M Protein Binds ADP. *Mol. Plant-Microbe Interact.* **24**, 897–906 (2011).
306. Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A Modular Cloning System for Standardized Assembly of Multigene Constructs. *PLoS One* **6**, e16765 (2011).
307. Hatta, M. A. M. *et al.* The wheat Sr22, Sr33, Sr35 and Sr45 genes confer resistance against stem rust in barley. *bioRxiv* 374637 (2018). doi:10.1101/374637
308. Patron, N. J. *et al.* Standards for plant synthetic biology: a common syntax for exchange of DNA parts. *New Phytol.* **208**, 13–19 (2015).
309. Lazo, G. R., Stein, P. A. & Ludwig, R. A. A DNA transformation-competent Arabidopsis genomic library in Agrobacterium. *Biotechnology. (N. Y).* **9**, 963–7 (1991).
310. Rey, M.-D. *et al.* Magnesium Increases Homoeologous Crossover Frequency During Meiosis in ZIP4 (Ph1 Gene) Mutant Wheat-Wild Relative Hybrids. *Front. Plant Sci.* **9**, 509 (2018).
311. Murashige, T. & Skoog, F. A Revised Medium for Rapid Growth and Bio Assays

with Tobacco Tissue Cultures. *Physiol. Plant.* **15**, 473–497 (1962).
312. Harwood*, W. A. *et al.* Barley Transformation Using Agrobacterium-Mediated Techniques. in 137–147 (Humana Press, 2009). doi:10.1007/978-1-59745-379-0_9
313. Deslandes, L. *et al.* Physical interaction between RRS1-R, a protein conferring resistance to bacterial wilt, and PopP2, a type III effector targeted to the plant nucleus. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8024–9 (2003).
314. Ghosh, S. *et al.* Speed breeding in growth chambers and glasshouses for crop breeding and model plant research. *Nat. Protoc.* **13**, 2944–2963 (2018).
315. Watson, A. *et al.* Speed breeding is a powerful tool to accelerate crop research and breeding. *Nat. Plants* **4**, 23–29 (2018).
316. Zuluaga, P., Szurek, B., Koebnik, R., Kroj, T. & Morel, J.-B. Effector mimics and integrated decoys, the never-ending arms race between rice and *Xanthomonas oryzae*. *Front. Plant Sci.* **8**, 431 (2017).
317. Wulff, B. B. & Moscou, M. J. Strategies for transferring resistance into wheat: from wide crosses to GM cassettes. *Front Plant Sci* **5**, 692 (2014).
318. Mundt, C. C. Pyramiding for Resistance Durability: Theory and Practice. *Phytopathology* **108**, 792–802 (2018).
319. Dong, O. X. & Ronald, P. C. Genetic Engineering for Disease Resistance in Plants: Recent Progress and Future Perspectives. *Plant Physiol.* **180**, 26–38 (2019).
320. Fuchs, M. Pyramiding resistance-conferring gene sequences in crops. *Curr. Opin. Virol.* **26**, 36–42 (2017).
321. McIntosh, R. A. *et al.* Rye-derived powdery mildew resistance gene Pm8 in wheat is suppressed by the Pm3 locus. *Theor. Appl. Genet.* **123**, 359–367 (2011).
322. Que, Q. *et al.* Trait stacking in transgenic crops: Challenges and opportunities. *GM Crops* **1**, 220–229 (2010).
323. Ainley, W. M. *et al.* Trait stacking via targeted genome editing. *Plant Biotechnol. J.* **11**, 1126–1134 (2013).
324. Guo, L. *et al.* Specific recognition of two MAX effectors by integrated HMA domains in plant immune receptors involves distinct binding surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 11637–11642 (2018).

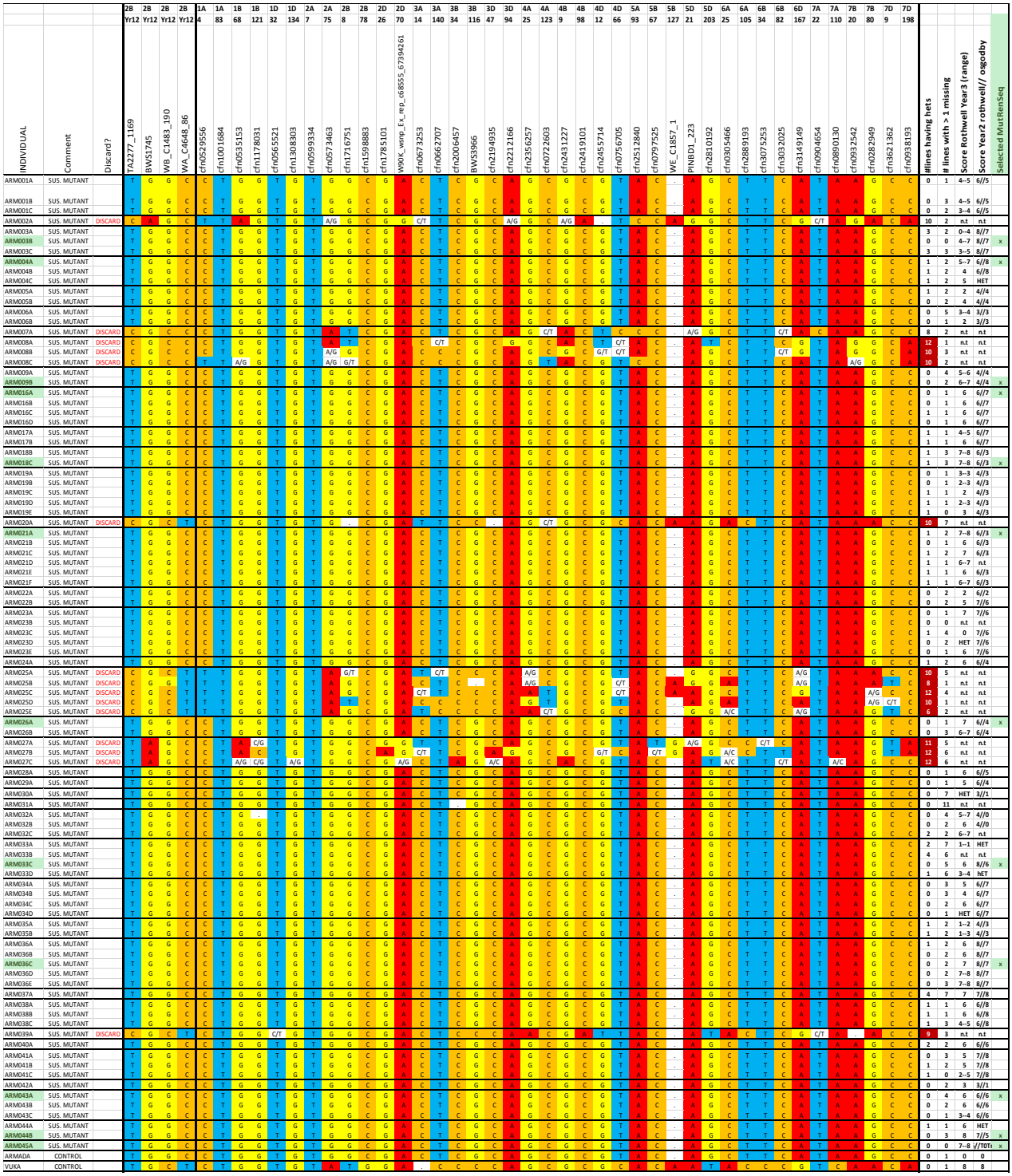
8. Appendices

Appendix 8-1. Harvested weight of known *Yr7* cultivars from 1990 to 2016 and prevalence of *Yr7* virulence among UK *Pst* isolates.
Proportion of harvested *Yr7* wheat varieties in the UK from 1990 to 2016. The prevalence of yellow rust isolates virulent to *Yr7* across this time period is shown in the top row. Original data from NIAB-TAG Seedstats journal (NIAB-TAG Network) and the UK Cereal Pathogen Virulence Survey (http://www.niab.com/pages/id/316/UKCPVS). See Figure 2-1 for an illustration of the data. Data published in Marchal et al., 2018

Cultivated Yr7 varieties		1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
	%virYr7_isolat	9	19	7	8	4	0	3	7	4	10	4	0	3	36	4	8	11	4	0	0	24	70	97	92	93	76	92
CORDIALE	total tons	0	0	0	0	0	0	0	0	0	0	0	0	21	969	5307	4819	6466	8013	10764	12346	10494	9171	8389	6,815.20	6,375.10	4,858.90	3,076.30
	%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	2.9	3.1	4.3	4.3	5.7	7.1	5.7	4.7	4.9	4.0	3.9	2.8	1.9
CUBANITA	total tons	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	65.9	490.9	197.7	53.9
	%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.0
GRAFTON	total tons	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	191	5010	10719	9948	9832	8,161.10	5,903.30	4,664.20	3,326.20
	%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	2.9	5.8	5.0	5.7	4.8	3.6	2.7	2.1
SKYFALL	total tons	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	275	11,885.60	17,032.90	17,587.70
	%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	7.2	9.7	11.0
RUSKIN	total tons	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13.8	9.20	0	0
	%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BROCK	total tons	3666.8	934.4	389	127.3	80.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	%	1.3	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CADENZA	total tons	0	0	337.5	8011.3	8412.3	3345.3	1146.4	634.5	744.8	223.5	234.8	132.65	117	60	39	0	0	0	0	0	0	0	0	0	0	0	0
	%	0.0	0.0	0.1	3.1	3.4	1.3	0.4	0.3	0.3	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CAMP REMY	total tons	1450.35	462.7	217	215.9	81.7	56.8	31.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	%	0.5	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PROPHET	total tons	0	0	0	124.2	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SOLEIL	total tons	65	47.7	152.5	71.5	60	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	%	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SPARK	total tons	0	0	2402.7	3734.2	3240.6	2737.9	2369.6	1627.1	1036.9	809.3	896.9	259.544	212.345	195	79	139	33	1	1	0	0	0	0	0	0	0	0
	%	0.0	0.0	1.0	1.5	1.3	1.0	0.9	0.7	0.5	0.4	0.5	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TARA	total tons	392.3	3018.7	748	85.7	49.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	%	0.1	1.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	total varieties	282286	283787	240546	255647	245240	261883	270400	247852	229351	222203	182648	176431	165486	186474	185970	154906	151525	184903	188184	174779	184795	197221	171034	170,276.70	164,779.00	174,991.40	159,371.70
	total %Yr7	2.0	1.6	1.8	4.8	4.9	2.4	1.3	0.9	0.8	0.5	0.6	0.2	0.2	0.7	2.9	3.2	4.3	4.3	5.8	9.9	11.5	9.7	10.7	9.0	15.0	15.3	15.1

Appendix 8-2. Summary of genotyping and phenotyping of M_3 and M_4 plants evaluated in the field for *Yr12* loss of function.

ARM001A refers to the M_1 lines ARM001, M_2 spike A. Armada wild-type and Vuka controls are shown at the bottom of the table. From left to right: Line identifier as described above, control or susceptible line, discarded or not based on genotype, genotype data for the 48 tested markers, number of heterozygotes lines, number of missing data, score in Year3 (range from the two to three tested progenies per original spike), score in Year2 in Rothwell and Osgodby, Selected for MutRenSeq. Note: breakdown of the number of progenies genotyped and phenotyped per original spike is not shown here for readability.



INDIVIDUAL	Comment	Discard?	TA2277_1169	BW51745	WB_C1483_190	WA_C4648_86	cfn0529556	cfn1001684	cfn0535153	cfn1178031	cfn0565521	cfn1308303	cfn0599334	cfn0573463	cfn1716751	cfn1598883	cfn1785101	W90k_wenp_Ex_rep_c68555_67394261	cfn0673253	cfn0662707	cfn2006457	BW53966	cfn2194935	cfn2212166	cfn2356257	cfn0722603	cfn2431227	cfn2419101	cfn2455714	cfn0756705	cfn2512840	cfn0797525	WE_C1857_1	PINBD1_223	cfn2810192	cfn0305466	cfn2889193	cfn3075253	cfn3032025	cfn3149149	cfn0904654	cfn0890130	cfn0932542	cfn0282949	cfn3621362	cfn0938193	#lines having hets	# lines with > 1 missing	Score Rothwell Year3 (range)	Score Year2 rothwell// osgodby	Selected MutRenSeq
ARM001A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	4–5	6//5	
ARM001B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	3	4–5	6//5	
ARM001C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	3–4	6//5	
ARM002A	SUS. MUTANT	DISCARD	C	A	G	C	T	T	A	G	T	G	T	A/G	G	C	G	G	C/T	T	C	G	C	A/G	G	C	A/G	A	.	T	C	C	A	G	G	C	T	T	C	G	C/T	A	G	A	C	A	10	2	nt	nt	
ARM003A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	3	2	0–4	8//7	
ARM003B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	0	4–7	8//7	x
ARM003C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	3	3	3–5	8//7	
ARM004A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	5–7	6//8	x
ARM004B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	4	6//8	
ARM004C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	5	HET	
ARM005A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	2	4//4	
ARM005B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	4	4//4	
ARM006A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	5	3–4	3//3	
ARM006B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	2	3//3	
ARM007A	SUS. MUTANT	DISCARD	C	G	C	C	C	T	G	G	T	G	T	T	C	G	A	C	T	C	G	C	A	G	C/T	A	C	T	C	C	.	A/G	G	C	T	T	C/T	A	C	A	G	C	C	A	G	C	8	2	nt	nt	
ARM008A	SUS. MUTANT	DISCARD	C	G	C	C	C	T	G	G	T	G	T	A	T	C	G	A	C	C/T	C	G	C	G	G	C	A	C	T	C/T	A	C	.	A	T	C	T	T	C	G	T	A	G	G	C	A	12	1	nt	nt	
ARM008B	SUS. MUTANT	DISCARD	C	G	C	C	C	T	G	G	T	G	T	A/G	G	C	G	C	C	C	G	C	A	G	C	G	G/T	C/T	A	C	.	A	G	C	T	T	C/T	G	T	A	G	G	C	A	10	3	nt	nt	
ARM008C	SUS. MUTANT	DISCARD	C	G	C	C	T	T	A/G	G	T	G	T	A/G	G/T	C	G	A	C	C	C	G	C	A	G	T	A	C	G	T	C	C	.	A	G	C	T	T	C	A	T	A	A/G	G	C	A	10	2	nt	nt	
ARM009A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	4	5–6	4//4	
ARM009B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	6–7	4//4	x
ARM016A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	6	6//7	x
ARM016B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	6	6//7	
ARM016C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6	6//7	
ARM016D	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	6	6//7	
ARM017A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	4–5	6//7	
ARM017B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6	6//7	
ARM018B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	3	7–8	6//3	
ARM018C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	3	7–8	6//3	x
ARM019A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	3–3	4//3	
ARM019B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	2–3	4//3	
ARM019C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	2	4//3	
ARM019D	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	2–3	4//3	
ARM019E	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	0	3	4//3	
ARM020A	SUS. MUTANT	DISCARD	C	G	C	T	C	T	G	G	T	G	T	G	.	C	G	A	T	T	C	C	.	A	G	C/T	G	C	G	C	A	C	A	A	G	A	C	T	T	C	A	T	A	A	A	C	C	10	7	nt	nt	
ARM021A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	7–8	6//3	x
ARM021B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	6	6//3	
ARM021C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	7	6//3	
ARM021D	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6–7	nt	
ARM021E	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6	6//3	
ARM021F	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6–7	6//3	
ARM022A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	2	6//2	
ARM022B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	5	7//6	
ARM023A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	7	7//6	
ARM023B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	0	nt	nt	
ARM023C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	4	0	7//6	
ARM023D	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	HET	7//6	
ARM023E	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	6	7//6	
ARM024A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	6	6//4	
ARM025A	SUS. MUTANT	DISCARD	C	G	C	T	T	T	G	G	T	G	T	A	G/T	C	G	A	T	C/T	C	C	C	A	A/G	C	G	C	G	T	A	C	.	G	G	C	T	T	C	A/G	T	A	A	A	C	C	10	5	nt	nt	
ARM025B	SUS. MUTANT	DISCARD	C	G	G	T	T	T	G	G	T	G	T	A	G	C	G	A	C	T	C	.	C	A	A/G	C	G	C	G	C/T	A	C	A	G	G	A	T	T	C	A/G	T	A	A	A	T	C	8	1	nt	nt	
ARM025C	SUS. MUTANT	DISCARD	C	G	C	T	T	T	G	G	T	G	T	A	G	C	G	A	C/T	T	C	C	C	A	A	T	G	C	G	C/T	A	C	A	A	G	C	T	T	C	G	T	A	A	A/G	C	C	12	4	nt	nt	
ARM025D	SUS. MUTANT	DISCARD	C	G	C	T	T	T	G	G	T	G	T	A	T	C	G	A	C	C	C	C	C	A	G	T	G	C	G	C	A	C	A	A	G	A	T	T	C	A/G	C/T	C	C	A	C	C	10	1	nt	nt	
ARM025E	SUS. MUTANT	DISCARD	C	G	C	T	T	T	G	G	T	G	T	A	T	C	G	A	T	C	C	C	C	A	A	C/T	G	C	G	C	A	C	A	A	G	A/C	T	T	C	A/G	T	A	A	G	C	C	6	2	nt	nt	
ARM026A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	7	6//4	x
ARM026B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	3	6–7	6//4	
ARM027A	SUS. MUTANT	DISCARD	T	A	G	C	C	T	A	C/G	T	G	T	G	G	C	G	G	T	T	C	G	C	A	G	C	G	C	G	T	A	T	G	A/G	G	C	C	C/T	C	A	T	A	A	G	T	A	11	5	nt	nt	
ARM027B	SUS. MUTANT	DISCARD	T	A	G	C	C	T	A	C	T	G	G	G	C	G	A	C/T	T	C	G	A	G	G	C	G	C	A	G/T	C	A	C/T	G	A	A/C	C	T	T	A	T	A	A	G	T	A	12	6	nt	nt	
ARM027C	SUS. MUTANT	DISCARD	T	A	G	C	C	T	A/G	C/G	T	A/G	T	G	G	C	G	A/G	C	T	A	G	A/C	A	G	C	A	C	G	T	A	C	.	A	T	A/C	T	T	C/T	A	T	A/C	A	G	C	C	12	6	nt	nt	
ARM028A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	6	6//5	
ARM029A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	5	6//4	
ARM030A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	7	HET	3//1	
ARM031A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	.	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	11	nt	nt	
ARM032A	SUS. MUTANT		T	G	G	C	C	T	G	G	.	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	4	5–7	4//0	
ARM032B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	6	4//0	
ARM032C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	2	2	6–7	nt	
ARM033A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	2	7	1–1	HET	
ARM033B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	4	6	nt	nt	
ARM033C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	5	6	8//6	x
ARM033D	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	6	3–4	HET	
ARM034A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	3	5	6//7	
ARM034B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	3	4	6//7	
ARM034C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	6	6//7	
ARM034D	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	HET	6//7	
ARM035A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	1–2	4//3	
ARM035B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	1–3	4//3	
ARM036A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	6	8//7	
ARM036B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	6	8//7	
ARM036C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	7	8//7	x
ARM036D	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	7–8	8//7	
ARM036E	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	3	7–8	8//7	
ARM037A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	4	7	7	7//8	
ARM038A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6	6//8	
ARM038B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6	6//8	
ARM038C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	3	4–5	6//8	
ARM039A	SUS. MUTANT	DISCARD	C	G	C	T	C	T	G	G	C/T	G	T	G	G	C	G	A	C	T	C	C	C	A	A	C	G	A	T	T	A	C	.	A	T	A	C	T	C	G	C/T	A	.	A	C	C	9	3	nt	nt	
ARM040A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	2	2	6	6//6	
ARM041A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	3	5	7//8	
ARM041B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	2	5	7//8	
ARM041C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	0	2–5	7//8	
ARM042A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	3	3//1	
ARM043A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	4	6	6//6	x
ARM043B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	2	6	6//6	
ARM043C	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	3–4	6//6	
ARM044A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	1	1	6	HET	
ARM044B	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	3	8	7//5	x
ARM045A	SUS. MUTANT		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	0	7–8	//TDT	x
ARMADA	CONTROL		T	G	G	C	C	T	G	G	T	G	T	G	G	C	G	A	C	T	C	G	C	A	G	C	G	C	G	T	A	C	.	A	G	C	T	T	C	A	T	A	A	G	C	C	0	1	0	0	
VUKA	CONTROL		T	G	C	T	C	T	G	G	T	G	T	A	T	G	G	A	.	C	C	C	C	A	G	C	G	C	G	C	A	C	A	A	T	A	C	C	C	G	T	C	A	A	C	A	0	1	8	8	

Appendix 8-3. Plant material submitted for Resistance gene enrichment Sequencing (RenSeq).
From left to right: Mutant line identifier, targeted gene, infection type (IT) when infected with *Pst* according to the Grassner and Straib scale.
*: Score from McGrann et al.

Accession	Target	IT
Cadenza-WT	*Yr7*	0nc
Cad127	*Yr7*	1
Cad1551	*Yr7*	1+
Cad1978	*Yr7*	2
Cad903	*Yr7*	1+
Cad923	*Yr7*	2
Cad855	*Yr7*	2
Cad1034	*Yr7*	2
AvocetS-Yr7	*Yr7*	0nc
AvSYr7_1	*Yr7*	3+
AvSYr7_2	*Yr7*	3+
AvSYr7_3	*Yr7*	3+
Lemhi-Yr5	*Yr5*	0;
Lem095	*Yr5*	4*
Lem387	*Yr5*	4*
Lem287	*Yr5*	4*
Lem241	*Yr5*	4*
Lem500	*Yr5*	4*
Lem115	*Yr5*	4*
Lem474	*Yr5*	4*
AvocetS-Yr5	*Yr5*	0;
AvSYr5_1	*Yr5*	3+
AvSYr5_2	*Yr5*	3+
AvSYr5_3	*Yr5*	3+
AvSYr5_4†	*Yr5*	3+
AvocetS+YrSP	*YrSP*	0
AvSYrSP_1	*YrSP*	3+
AvSYrSP_2	*YrSP*	3+
AvSYrSP_3	*YrSP*	3+
AvSYrSP_4	*YrSP*	3+

Appendix 8-4. Plant material submitted for Resistance gene enrichment Sequencing (RenSeq).
From left to right: Mutant line identifier, targeted gene, infection type (IT) when infected with PST according to the Grassner and Straib scale, mutation position, coverage of the mutation (at least 99 % of the reads supported the mutant base in the mutant reads), predicted effect of the mutation on the protein sequence. Lines with the same mutations are highlighted with matching colours. *: Score from McGrann et al., 2014; †: Used for Sanger sequencing confirmation only

Accession	Target	IT	Mutation	Coverage (X)	Predicted Effect on Protein
Cadenza-WT	*Yr7*	0nc	-	-	-
Cad127	*Yr7*	1	G4364A	46	W1352*
Cad1551	*Yr7*	1+	C3193T	193	T962M
Cad1978	*Yr7*	2	C2994T	312	H896Y
Cad903	*Yr7*	1+	G609A	50	G173R
Cad923	*Yr7*	2	C2917T	338	S870F
Cad855	*Yr7*	2	C3231T	311	P975S
Cad1034	*Yr7*	2	G2113A	338	G602E
AvocetS-Yr7	*Yr7*	0nc	-	-	-
AvSYr7_1	*Yr7*	3+	G2910A		G868R
AvSYr7_2	*Yr7*	3+	C1606T		P433L
AvSYr7_3	*Yr7*	3+	C1606T		P433L
Lemhi-Yr5	*Yr5*	0;	-	-	-
Lem095	*Yr5*	4*	G1680A	525	W481*
Lem387	*Yr5*	4*	G718A	47	Splice junction
Lem287	*Yr5*	4*	C4159T	36	Q1308*
Lem241	*Yr5*	4*	G718A	38	Splice junction
Lem500	*Yr5*	4*	G2901A	321	W888*
Lem115	*Yr5*	4*	C2260T	395	H675Y
Lem474	*Yr5*	4*	C1924T	247	L563F
AvocetS-Yr5	*Yr5*	0;	-	-	-
AvSYr5_1	*Yr5*	3+	G1237A	176	G334S
AvSYr5_2	*Yr5*	3+	G3475A	237	D1080N
AvSYr5_3	*Yr5*	3+	C2914T	236	L893F
AvSYr5_4†	*Yr5*	3+	G1748A	-	G504E
AvocetS+YrSP	*YrSP*	0	-	-	-
AvSYrSP_1	*YrSP*	3+	C2246T	114	P670L
AvSYrSP_2	*YrSP*	3+	G2068A	525	G611R
AvSYrSP_3	*YrSP*	3+	C2317T	352	R694C
AvSYrSP_4	*YrSP*	3+	C2476T	355	P747S
Paragon	*Yr7*	0nc	-	-	-
Kronos	*Yr5*		-	-	-

Appendix 8-5. Sequencing details of RenSeq data generated in this study.

Sample	Accession	SRA identifier	Sequencing chenistry	Enrichment pool	Sequence pool	#Read-pairs	#Read-pairs after trimming	#Read-pairs mapped to the de novo assembly	%Read-pairs mapped to the *de novo* assembly	*de novo* assembly
MW01-127_HM7MVBCXX_L1_1.fq.gz	Cad127	SAMN08897506	Illumina_HiSeq_2500 (250bp PE)	A	1	14805176	14743094	18772686	64%	Cadenza-WT
MW01-127_HM7MVBCXX_L1_2.fq.gz	Cad127	SAMN08897506	Illumina_HiSeq_2500 (250bp PE)	A	1	14805176	14743094			
MW01-1551_HM7MVBCXX_L1_1.fq.gz	Cad1551	SAMN08897507	Illumina_HiSeq_2500 (250bp PE)	A	1	8216218	8184048	10619188	65%	Cadenza-WT
MW01-1551_HM7MVBCXX_L1_2.fq.gz	Cad1551	SAMN08897507	Illumina_HiSeq_2500 (250bp PE)	A	1	8216218	8184048			
MW01-1978_HM7MVBCXX_L1_1.fq.gz	Cad1978	SAMN08897508	Illumina_HiSeq_2500 (250bp PE)	B	1	12462294	12409066	15916836	64%	Cadenza-WT
MW01-1978_HM7MVBCXX_L1_2.fq.gz	Cad1978	SAMN08897508	Illumina_HiSeq_2500 (250bp PE)	B	1	12462294	12409066			
WW01-27_Cadenza_S3_L001_R1_001.fastq.gz	Cadenza-WT	SAMN08897509	Illumina_MiSeq (250bp PE)	C	2	5901019	5843683	7884202	67%	Cadenza-WT
WW01-27_Cadenza_S3_L001_R2_001.fastq.gz	Cadenza-WT	SAMN08897509	Illumina_MiSeq (250bp PE)	C	2	5901019	5843683			
AvS_KD17010810-A71_HCHT7BCXY_L1_1.fq.gz	AvocetS	SAMN08897510	Illumina_HiSeq_2500 (250bp PE)	D	3	12669666	12284950	NA	NA	used to confirm the presence of the candidates in the corresponding Near Isogenic Line
AvS_KD17010810-A71_HCHT7BCXY_L1_2.fq.gz	AvocetS	SAMN08897510	Illumina_HiSeq_2500 (250bp PE)	D	3	12669666	12284950			
AvS_SP_KD17010810-A50_HCHT7BCXY_L1_1.fq.gz	AvocetS-YrSP	SAMN08897511	Illumina_HiSeq_2500 (250bp PE)	D	3	13559810	12937267	NA	NA	used to confirm the presence of the candidates in the corresponding Near Isogenic Line
AvS_SP_KD17010810-A50_HCHT7BCXY_L1_2.fq.gz	AvocetS-YrSP	SAMN08897511	Illumina_HiSeq_2500 (250bp PE)	D	3	13559810	12937267			
AvS_Yr5_KD17010810-A81_HCHT7BCXY_L1_1.fq.gz	AvocetS-Yr5	SAMN08897512	Illumina_HiSeq_2500 (250bp PE)	D	3	10131809	9734508	NA	NA	used to confirm the presence of the candidates in the

Sample	Accession	SRA identifier	Sequencing chenistry	Enrichment pool	Sequence pool	#Read-pairs	#Read-pairs after trimming	#Read-pairs mapped to the de novo assembly	%Read-pairs mapped to the *de novo* assembly	*de novo* assembly
										corresponding Near Isogenic Line
AvS_Yr5_KD17010810-A81_HCHT7BCXY_L1_2.fq.gz	AvocetS-Yr5	SAMN08897512	Illumina_HiSeq_2500 (250bp PE)	D	3	10131809	9734508			
AvS_Yr7_KD17010810-A93_HCHT7BCXY_L1_1.fq.gz	AvocetS-Yr7	SAMN08897513	Illumina_HiSeq_2500 (250bp PE)	D	3	7698058	7244558	NA	NA	used to confirm the presence of the candidates in the corresponding Near Isogenic Line
AvS_Yr7_KD17010810-A93_HCHT7BCXY_L1_2.fq.gz	AvocetS-Yr7	SAMN08897513	Illumina_HiSeq_2500 (250bp PE)	D	3	7698058	7244558			
C855_KD17010810-A2_HCHT7BCXY_L1_1.fq.gz	Cad855	SAMN08897514	Illumina_HiSeq_2500 (250bp PE)	E	3	13109055	12568140	17166458	68%	Cadenza-WT
C855_KD17010810-A2_HCHT7BCXY_L1_2.fq.gz	Cad855	SAMN08897514	Illumina_HiSeq_2500 (250bp PE)	E	3	13109055	12568140			
C903_KD17010810-A94_HCHT7BCXY_L1_1.fq.gz	Cad903	SAMN08897515	Illumina_HiSeq_2500 (250bp PE)	E	3	9109264	8704600	11780688	68%	Cadenza-WT
C903_KD17010810-A94_HCHT7BCXY_L1_2.fq.gz	Cad903	SAMN08897515	Illumina_HiSeq_2500 (250bp PE)	E	3	9109264	8704600			
C923_KD17010810-A40_HCHT7BCXY_L1_1.fq.gz	Cad923	SAMN08897516	Illumina_HiSeq_2500 (250bp PE)	E	3	14252713	13647531	17530654	64%	Cadenza-WT
C923_KD17010810-A40_HCHT7BCXY_L1_2.fq.gz	Cad923	SAMN08897516	Illumina_HiSeq_2500 (250bp PE)	E	3	14252713	13647531			
C1034_KD17010810-A49_HCHT7BCXY_L1_1.fq.gz	Cad1034	SAMN08897517	Illumina_HiSeq_2500 (250bp PE)	E	3	13415313	12889224	15567764	60%	Cadenza-WT
C1034_KD17010810-A49_HCHT7BCXY_L1_2.fq.gz	Cad1034	SAMN08897517	Illumina_HiSeq_2500 (250bp PE)	E	3	13415313	12889224			
YSP_0_KD17071213-AK3122_HV32GBCXY_L1_1.fq.gz	AvocetS-YrSP-WT	SAMN09091012	Illumina_HiSeq_2500 (250bp PE)	F	4	20168141	19285244	25472610	66%	AvocetS-YrSP-WT
YSP_0_KD17071213-AK3122_HV32GBCXY_L1_2.fq.gz	AvocetS-YrSP-WT	SAMN09091012	Illumina_HiSeq_2500 (250bp PE)	F	4	20168141	19285244			AvocetS-YrSP-WT
YSP_1_KD17071213-AK2489_HV32GBCXY_L1_1.fq.gz	AvocetS-YrSP-M1	SAMN09091013	Illumina_HiSeq_2500 (250bp PE)	F	4	4866592	4715938	6208114	66%	AvocetS-YrSP-WT
YSP_1_KD17071213-AK2489_HV32GBCXY_L1_2.fq.gz	AvocetS-YrSP-M1	SAMN09091013	Illumina_HiSeq_2500 (250bp PE)	F	4	4866592	4715938			AvocetS-YrSP-WT

Sample	Accession	SRA identifier	Sequencing chenistry	Enrichment pool	Sequence pool	#Read-pairs	#Read-pairs after trimming	#Read-pairs mapped to the de novo assembly	%Read-pairs mapped to the *de novo* assembly	*de novo* assembly
YSP_2_KD17071213-AK3121_HV32GBCXY_L1_1.fq.gz	AvocetS-YrSP-M2	SAMN09091014	Illumina_HiSeq_2500 (250bp PE)	G	4	22067358	21281452	28040118	66%	AvocetS-YrSP-WT
YSP_2_KD17071213-AK3121_HV32GBCXY_L1_2.fq.gz	AvocetS-YrSP-M2	SAMN09091014	Illumina_HiSeq_2500 (250bp PE)	G	4	22067358	21281452			AvocetS-YrSP-WT
YSP_3_KD17071213-AK2464_HV32GBCXY_L1_1.fq.gz	AvocetS-YrSP-M3	SAMN09091015	Illumina_HiSeq_2500 (250bp PE)	G	4	14603831	14068492	18132636	64%	AvocetS-YrSP-WT
YSP_3_KD17071213-AK2464_HV32GBCXY_L1_2.fq.gz	AvocetS-YrSP-M3	SAMN09091015	Illumina_HiSeq_2500 (250bp PE)	G	4	14603831	14068492			AvocetS-YrSP-WT
YSP_4_KD17071213-AK2483_HV32GBCXY_L1_1.fq.gz	AvocetS-YrSP-M4	SAMN09091016	Illumina_HiSeq_2500 (250bp PE)	H	4	16757582	15993630	20438956	64%	AvocetS-YrSP-WT
YSP_4_KD17071213-AK2483_HV32GBCXY_L1_2.fq.gz	AvocetS-YrSP-M4	SAMN09091016	Illumina_HiSeq_2500 (250bp PE)	H	4	16757582	15993630			AvocetS-YrSP-WT
Y5_0_KD17071213-AK2488_HV32GBCXY_L1_1.fq.gz	AvocetS-Yr5-WT	SAMN09091017	Illumina_HiSeq_2500 (250bp PE)	H	4	18106714	17329780	23756414	69%	AvocetS-Yr5-WT
Y5_0_KD17071213-AK2488_HV32GBCXY_L1_2.fq.gz	AvocetS-Yr5-WT	SAMN09091017	Illumina_HiSeq_2500 (250bp PE)	H	4	18106714	17329780			AvocetS-Yr5-WT
Y5_1_KD17071213-AK2485_HV32GBCXY_L1_1.fq.gz	AvocetS-Yr5-M1	SAMN09091018	Illumina_HiSeq_2500 (250bp PE)	I	4	12149902	11617256	14917602	64%	AvocetS-Yr5-WT
Y5_1_KD17071213-AK2485_HV32GBCXY_L1_2.fq.gz	AvocetS-Yr5-M1	SAMN09091018	Illumina_HiSeq_2500 (250bp PE)	I	4	12149902	11617256			AvocetS-Yr5-WT
Y5_2_KD17071213-AK2486_HV32GBCXY_L1_1.fq.gz	AvocetS-Yr5-M2	SAMN09091019	Illumina_HiSeq_2500 (250bp PE)	I	4	18064931	16987606	23153166	68%	AvocetS-Yr5-WT
Y5_2_KD17071213-AK2486_HV32GBCXY_L1_2.fq.gz	AvocetS-Yr5-M2	SAMN09091019	Illumina_HiSeq_2500 (250bp PE)	I	4	18064931	16987606			AvocetS-Yr5-WT
Y5_3_KD17071213-AK2487_HV32GBCXY_L1_1.fq.gz	AvocetS-Yr5-M3	SAMN09091020	Illumina_HiSeq_2500 (250bp PE)	J	4	15563606	14814817	19915922	67%	AvocetS-Yr5-WT
Y5_3_KD17071213-AK2487_HV32GBCXY_L1_2.fq.gz	AvocetS-Yr5-M3	SAMN09091020	Illumina_HiSeq_2500 (250bp PE)	J	4	15563606	14814817			AvocetS-Yr5-WT
Paragon_DO16074003-5_HNWN7BCXX_L1_1.clean.fq.gz	Paragon	SAMN08897526	Illumina_HiSeq_2500 (250bp PE)	K	5	20292064	NA	NA	NA	used to confirm the presence of *Yr7* in the *de novo* assembly
Paragon_DO16074003-5_HNWN7BCXX_L1_2.clean.fq.gz	Paragon	SAMN08897526	Illumina_HiSeq_2500 (250bp PE)	K	5	20292064	NA	NA	NA	used to confirm the presence of *Yr7* in the *de novo* assembly
WW01-26_Kronos_S2_L001_R1_001.fastq.gz	Kronos	SAMN08897527	Illumina_MiSeq (250bp PE)	L	6	5877285	NA	NA	NA	used to confirm the presence of the *Yr5*

Sample	Accession	SRA identifier	Sequencing chenistry	Enrichment pool	Sequence pool	#Read-pairs	#Read-pairs after trimming	#Read-pairs mapped to the de novo assembly	%Read-pairs mapped to the *de novo* assembly	*de novo* assembly
										alternate allele in the *de novo* assembly
WW01-26_Kronos_S2_L001_R2_001.fastq.gz	Kronos	SAMN08897527	Illumina_MiSeq (250bp PE)	L	6	5877285	NA	NA	NA	used to confirm the presence of the *Yr5* alternate allele in the *de novo* assembly
L115_KD17051870-AK2850_HMLMJBCXY_L1_1.fq.gz	Lem115	SAMN08897518	Illumina_HiSeq_2500 (250bp PE)	M	7	10222627	10222493	13935550	68%	Lemhi-Yr5
L115_KD17051870-AK2850_HMLMJBCXY_L1_2.fq.gz	Lem115	SAMN08897518	Illumina_HiSeq_2500 (250bp PE)	M	7	10222627	10222493			Lemhi-Yr5
L241_KD17051870-AK618_HMLMJBCXY_L1_1.fq.gz	Lem241	SAMN08897519	Illumina_HiSeq_2500 (250bp PE)	M	7	10482636	10482519	13131331	63%	Lemhi-Yr5
L241_KD17051870-AK618_HMLMJBCXY_L1_2.fq.gz	Lem241	SAMN08897519	Illumina_HiSeq_2500 (250bp PE)	M	7	10482636	10482519			Lemhi-Yr5
L287_KD17051870-AK619_HMLMJBCXY_L1_1.fq.gz	Lem287	SAMN08897520	Illumina_HiSeq_2500 (250bp PE)	M	7	13620621	13620411	17878002	66%	Lemhi-Yr5
L287_KD17051870-AK619_HMLMJBCXY_L1_2.fq.gz	Lem287	SAMN08897520	Illumina_HiSeq_2500 (250bp PE)	M	7	13620621	13620411			Lemhi-Yr5
L387_KD17051870-AK602_HMLMJBCXY_L1_1.fq.gz	Lem387	SAMN08897521	Illumina_HiSeq_2500 (250bp PE)	M	7	14803162	14802948	19969510	67%	Lemhi-Yr5
L387_KD17051870-AK602_HMLMJBCXY_L1_2.fq.gz	Lem387	SAMN08897521	Illumina_HiSeq_2500 (250bp PE)	M	7	14803162	14802948			Lemhi-Yr5
L474_KD17051870-AK361_HMLMJBCXY_L1_1.fq.gz	Lem474	SAMN08897522	Illumina_HiSeq_2500 (250bp PE)	N	7	13083995	13083778	19434030	74%	Lemhi-Yr5
L474_KD17051870-AK361_HMLMJBCXY_L1_2.fq.gz	Lem474	SAMN08897522	Illumina_HiSeq_2500 (250bp PE)	N	7	13083995	13083778			Lemhi-Yr5
L500_KD17051870-AK606_HMLMJBCXY_L1_1.fq.gz	Lem500	SAMN08897523	Illumina_HiSeq_2500 (250bp PE)	N	7	21430686	21430385	15380803	36%	Lemhi-Yr5
L500_KD17051870-AK606_HMLMJBCXY_L1_2.fq.gz	Lem500	SAMN08897523	Illumina_HiSeq_2500 (250bp PE)	N	7	21430686	21430385			Lemhi-Yr5
L592_KD17051870-AK2848_HMLMJBCXY_L1_1.fq.gz	Lem095	SAMN08897524	Illumina_HiSeq_2500 (250bp PE)	N	7	21885943	21885584	32007088	73%	Lemhi-Yr5
L592_KD17051870-AK2848_HMLMJBCXY_L1_2.fq.gz	Lem095	SAMN08897524	Illumina_HiSeq_2500 (250bp PE)	N	7	21885943	21885584			Lemhi-Yr5
LYr5_KD17051870-AK2849_HMLMJBCXY_L1_1.fq.gz	Lemhi-Yr5	SAMN08897525	Illumina_HiSeq_2500 (250bp PE)	N	7	11546335	11546185	13868897	60%	Lemhi-Yr5

Sample	Accession	SRA identifier	Sequencing chenistry	Enrichment pool	Sequence pool	#Read-pairs	#Read-pairs after trimming	#Read-pairs mapped to the de novo assembly	%Read-pairs mapped to the *de novo* assembly	*de novo* assembly
LYr5_KD17051870-AK2849_HMLMJBCXY_L1_2.fq.gz	Lemhi-Yr5	SAMN08897525	Illumina_HiSeq_2500 (250bp PE)	N	7	11546335	11546185			Lemhi-Yr5

Appendix 8-6. Summary of the available genome assemblies that were used for *in silico* allele mining

Specie	Cultivar/group	Source	Link/ref
Triticum aestivum	Chinese Spring	IWGSC	https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies
Triticum aestivum	Cadenza	EarIham Institute	http://opendata.earlham.ac.uk/Triticum_aestivum/EI/v1.1/
Triticum aestivum	Paragon	EarIham Institute	http://opendata.earlham.ac.uk/Triticum_aestivum/EI/v1.1/
Triticum aestivum	Claire	EarIham Institute	http://opendata.earlham.ac.uk/Triticum_aestivum/EI/v1.1/
Triticum aestivum	Robigus	EarIham Institute	http://opendata.earlham.ac.uk/Triticum_aestivum/EI/v1.1/
Triticum turgidum	Kronos	EarIham Institute	http://opendata.earlham.ac.uk/Triticum_turgidum/EI/v1.1/
Triticum turgidum	Svevo	The International Durum Wheat Genome Sequencing Consortium	http://d-data.interomics.eu
Triticum turgidum	Zavitan	WEWseq	Avni et al. 2017

Appendix 8-7. Primers designed to map and clone *Yr5*, *Yr7*, *YrSP*.
Note that KASP assays require the addition of the corresponding tails in the 5' for the two KASP primers.

Primer name	Gene	Primer type	chromosome	KASP wild-type allele	KASP mutant allele	common	product size (bp)
Cad127	Yr7	KASP	2BL	AAGTGATGTCGGGAGGAGc	AAGTGATGTCGGGAGGAGt	TGGAGAATGGAAGTTCTTTTGTGT	83
Cad1551	Yr7	KASP	2BL	CACAATCATCAAGATGAAGCg	CACAATCATCAAGATGAAGCa	CCAACAATATCTCAGTTACCTCATTG	51
Cad1978	Yr7	KASP	2BL	TGCATCCTTCCAGGACAAATg	TGCATCCTTCCAGGACAAATa	AACCAGGGAGGACGCTTATG	79
Cad127_M1	Yr7 mapping	KASP	2BL	ACATATTCGTGGAGGCCGg	ACATATTCGTGGAGGCCGa	TGGTGAACTCTGATAGGAACTTC	94
Cad127_M2	Yr7 mapping	KASP	2BL	TTCTCCTGCGCCTCTCTGg	TTCTCCTGCGCCTCTCTGa	GGAGGGTCTGGCCTCTGT	59
Cad127_M3	Yr7 mapping	KASP	2BL	CGGAACCAATCACCTCGGg	CGGAACCAATCACCTCGGa	ATGTTGTCCACGGCGATTAA	78
Cad127_M4	Yr7 mapping	KASP	2BL	GAAAGCAGCAGCCACAGc	GAAAGCAGCAGCCACAGt	TTGGTCGGCTCTTGAACTTT	55
Cad127_M5	Yr7 mapping	KASP	2BL	CATCATCCATTTTCCCTCTCGc	CATCATCCATTTTCCCTCTCGt	AGCTTCTTTAGAACATGCCAAC	51
Cad127_M6	Yr7 mapping	KASP	2BL	ACTGCTCGCAACACATACAc	ACTGCTCGCAACACATACAt	CCCAATTATTTGCAGTGCTTGAG	67
Cad127_M7	Yr7 mapping	KASP	2BL	GCTTCAGTGAACAAGGTGATGc	GCTTCAGTGAACAAGGTGATGt	GAGAGGAGAAATGACATCCTAGAT	36
Cad127_M8	Yr7 mapping	KASP	2BL	AGAACCAGAGAATTTGTTGTTGTAg	AGAACCAGAGAATTTGTTGTTGTAa	CGACTATGGAGAACCTTGAGAGA	103
Cad127_M9	Yr7 mapping	KASP	2BL	GCCTTTCTTCATCTGGCCTTTAGc	GCCTTTCTTCATCTGGCCTTTAGt	TGTGGTACGAGTTGGCATACC	78
Yr5_candidate	Yr5	KASP	2BL	CAGGAGATCTTGAAGGACAT	CAGGAGATCTTAAAGGAATA	AAACTCTTTGACTGGTACTCG	44
Yr5_M1	W90K_Kukri_c10138_391	KASP	2BL	ATATCACTGCTGCCTGTAGTGGA	ATCACTGCTGCCTGTAGTGGG	ACGAGTAGCTGTAATTAAACCAACAATGAA	53
Yr5_M2	W90K_RAC875_c29700_198	KASP	2BL	GGAATACCGCCTAGTAGATCAGT	GGAATACCGCCTAGTAGATCAGC	CGTCATAAACTCTTCACTCTTATGAGCTA	64
Yr5_M3	W90K_Tdurum_contig14707_185	KASP	2BL	AAGTTTACTTGGTTGGAGCATGGGA	GTTTACTTGGTTGGAGCATGGGG	GCCCAATAACACCGAAGGATGATCTT	62
Yr5_M4	W90K_Ra_c68109_376	KASP	2BL	ATCCTGGAGATGTGATGTGTGTTCA	CCTGGAGATGTGATGTGTGTTCG	GTCCTGGTGAAACAGGTCAAGATGAT	58
Yr5_M5	W90K_GENE-0675_104	KASP	2BL	ATGGTGTGCTTTTAAAGAATGCAGATATA	GGTGTGCTTTTAAAGAATGCAGATATG	TACACATTTGTGTAGAAGGTCGAGCAA	57
Yr5_M6	W90K_wsnp_Ex_c16425_24923685	KASP	2BL	CATCGGAGTCGACATCATCTTCA	CATCGGAGTCGACATCATCTTCG	GGGAGGCTGTAGAGTTGTCCCTA	51

Primer name	Gene	Primer type	chromosome	Forward	Reverse	product size (bp)	
WMC175		SSR	2BL	gcTcAgTcAAAccgcTAcTTcT	cAcTAcTccAATcTATcgccgT	253bp (Chinese Spring)	
Primer name	Gene	Primer type	chromosome	KASP target	KASP alternative base	common	product size (bp)
Yr5_candidate	YrSP	KASP	2BL	CAGGAGATCTTGAAGGACAT	CAGGAGATCTTAAAGGAATA	AAACTCTTTGACTGGTACTCG	44
YrSP_M1	W90K_JD_c2156_2040	KASP	2BL	GTGCTATTATTAGTAGTACTAAAATTTTGACT	GTGCTATTATTAGTAGTACTAAAATTTTGACC	GCATACGAGAATAATAATCTGCTGTCTGAA	66
YrSP_M2	RAC875_rep_c85788_282	KASP	2BL	ATCCCCAAGCAGCTCTGGGTTA	CCCCAAGCAGCTCTGGGTTG	CAGATTGTGCGCAAGAGGAATGTCAA	48
YrSP_M3	BobWhite_c3871_1170	KASP	2BL	CAGTTTTTCAAGCATGCCTTGGCTT	AGTTTTTCAAGCATGCCTTGGCTC	CACATCTTGTCGCCCTGGGGAA	51

Appendix 8-8. Diagnostic markers for *Yr5*, *Yr7*, *YrSP*.
Note that KASP assays require the addition of the corresponding tails in the 5' for the two allele primers (see comment)

Primer name	Gene	KASP_R-gene_allele	KASP_alternate_allele	common	product_size	Comment
Yr7-A	Yr7	TTAGTCCTGCCCCATAAGC**g**	TTAGTCCAGCCCCATAAGC**c**	CAGTGTTAAAACCAGGGAGGA	41	The SNP primers are reverse and the common one is forward
Yr7-B	Yr7	TGGAGGTATCATCTGGTGA**g**	TGGAGGTATCATCGGGTGA**a**	CATCAAAATCATCGCCTATGT	70	Dominant marker: alternate allele is actually not amplified
Yr7-D	Yr7	GCTGGAAAGGCTTGACATC**a**	GCTGGAAAGGCTTGAGATC**g**	AATGGCGTGGTAAGGACAGA	48	
YrSP	YrSP	GAGAAAATCAGCAGGTG**g**	GAGAAAATCAGCAGGTG**c**	AGCGAGTTGAGGACATTGGT	129	The SNP primers are reverse and the common one is forward
Yr5	Yr5	GCGCCCCTTTTCGAAAAAATA	CTAGCATCAAACAAGCTAAATA	ATGTCGAAATATTGCATAACATGG	83	See Figure 3-16

Appendix 8-9. Presence/absence of *Yr7* and *YrSP* in different wheat collections.
We used Vuka, AvocetS, and Solstice as negative controls for the presence of *Yr7* and *YrSP* and AvocetS-Yr near-isogenic lines as controls for the corresponding *Yr* gene. We genotyped different collections:
(i) a set of potential *Yr7* carriers based on literature research,
(ii) a set of varieties that belonged to the UK AHDB Recommended List (https://cereals.ahdb.org.uk/varieties/ahdb-recommended-lists.aspx) between 2005 and 2018 (labelled 2005-2018-UK_RL),
(iii) the Gediflux collection that includes modern European bread wheat varieties (1920-2010)[191],
(iv) a core set of the Watkins collection, which represent a set of global bread wheat landraces collected in the 1920-30s[190].
We separated the table in different parts according to the tested population to help with clarity.

				Yr7				*YrSP*	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	*Yr7*	*Yr5 / YrSP* KASP	*YrSP*
control	GBR	1992	Cadenza	C	G	A	*Yr7*	N-T	-
control			AvocetS-Yr7	C	G	A	*Yr7*	N-A	non-*YrSP*
control			AvocetS-YrSP	G	N-A	G	non-*Yr7*	C	*YrSP*
control			AvocetS-Yr5	G	N-A	G	non-*Yr7*	G	non-*YrSP*
control			AvocetS	G	N-A	G	non-*Yr7*	N-A	non-*YrSP*
control	HRV	1964	Vuka	G	N-A	G	non-*Yr7*	N-T	-
control	NLD	2002	Solstice	G	N-A	G	non-*Yr7*	G	non-*YrSP*
control			AvocetS-Yr7	C	G	A	*Yr7*	N-A	non-*YrSP*
control	GBR	1992	CADENZA	C	G	A	*Yr7*	G	non-*YrSP*
control			AvocetS	G	N-A	G	non-*Yr7*	N-A	non-*YrSP*
control			Lemhi-Yr5	G	N-A	G	non-*Yr7*	G	non-*YrSP*
potential *Yr7* carriers	FRA	1998	Apache	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	FRA	1994	Aztec	G	N-A	G	non-*Yr7*	N-T	-
potential *Yr7* carriers	GBR	1985	Brock	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	FRA	1980	Camp Remy	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	GBR	1994	Chablis	G	N-A	G	non-*Yr7*	N-T	-
potential *Yr7* carriers	GBR	2000	Cordiale	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	AUS:Western-Australi	1985	Cranbrook	G	N-A	G	non-*Yr7*	N-T	-
potential *Yr7* carriers	NLD	1979	Donata	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	NLD	1964	Flevina	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	RUS:Rostov	2005	Garant	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	FRA	1969	Hardi	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	USA:Minnesota	1950	Lee	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	NLD	1970	Lely	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	MEX	1976	Pavon 76	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	FRA	1978	Prinqual	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	GBR:England	1983	Renard	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	GBR	1991	Spark	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	FRA	1973	Talent	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	GBR	1987	Tara	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	USA:Minnesota	1934	Thatcher	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	FRA	1971	Tommy	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	GBR	1985	Tonic	C	G	A	*Yr7*	N-T	-
potential *Yr7* carriers	GBR	2002	Vector	C	G	A	*Yr7*	N-T	-

				Yr7				YrSP	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	Yr7	Yr5 /YrSP KASP	YrSP
2005-2018-UK_RL	GBR	2002	ACCESS	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2002	ALCHEMY	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2005	AMBROSIA	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2006	BANTAM	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2006	BATTALION	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2010	BELUGA	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2017	BENNINGTON	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1999	BISCAY	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1992	BRIGADIER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2015	BRITANNIA	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2005	BROMPTON	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1999	CANTERBURY	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2007	CASSIUS	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1992	CHARGER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	DEU	2011	CHILTON	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	1999	CLAIRE	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2010	COCOON	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2009	CONQUEROR	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR:England	1993	CONSORT	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2000	CORDIALE	C	G	A	Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2012	CRUSOE	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2015	CUBANITA	C	G	A	Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2000	DEBEN	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2010	DENMAN	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	FRA	2014	DICKENS	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	NLD	2003	DICKSON	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2017	DUNSTON	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2007	DUXFORD	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2007	EINSTEIN	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2014	ENERGISE	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1992	EQUINOX	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	DNK	2010	EVOLUTION	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	1995	FLAME	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2017	FREISTON	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2010	GALLANT	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2007	GATSBY	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	FRA	2006	GLASGOW	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2012	GRAFTON	C	G	A	Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2010	GRAVITAS	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	DNK	2017	HARDWICKE	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	DNK	2007	HEREFORD	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1989	HEREWARD	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2012	HORATIO	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2006	HUMBER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	FRA	2013	ICON	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2010	INVICTA	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2004	ISTABRAQ	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	DEU	2007	JB_DIEGO	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2014	JORVIK	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2005	KETCHUM	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2011	KINGDOM	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR		KWS_BARREL	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL			KWS_BASSET	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2014	KWS_CASHEL	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR		KWS_CRISPIN	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2010	KWS_CROFT	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2011	KWS_GATOR	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL			KWS_KERRIN	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2013	KWS_KIELDER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2015	KWS_LILI	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2010	KWS_PODIUM	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2010	KWS_SANTIAGO	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL			KWS_SILVERSTO	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR		KWS_SISKIN	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2015	KWS_TEMPO	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2007	LEAR	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	FRA	2010	LEEDS	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2016	LG MOTOWN	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2016	LG SUNDANCE	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1997	MALACCA	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2007	MARKSMAN	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2005	MASCOT	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2017	MOULTON	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR		MYRIAD	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1999	NAPIER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2009	NIJINSKY	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2008	OAKLEY	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2001	OPTION	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	FRA	2014	PANACEA	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	FRA	2007	PANORAMA	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2007	QPLUS	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	1996	REAPER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2011	RELAY	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2014	REVELATION	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	FRA	2016	RGT_ILLUSTRIOU	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2014	RGT_SCRUMMA	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR:England	1993	RIALTO	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR:England	1987	RIBAND	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	1998	RICHMOND	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2005	ROBIGUS	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	AUS:New-South-Wales	1990	RUSKIN	C	G	A	Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	1998	SAVANNAH	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2017	SAVELLO	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2008	SCOUT	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2017	SHABRAS	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	NLD	2006	SHEPHERD	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2015	SKYFALL	C	G	A	Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2002	SMUGGLER	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	FRA	1987	SOISSONS	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	NLD	2002	SOLSTICE	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2010	STIGG	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2002	TANKER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2008	TIMARU	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	FRA	2006	TIMBER	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	FRA	2011	TORCH	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR:England	1991	VIVANT	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	1990	WASP	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2006	WELFORD	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR:England	1976	WIZARD	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	NLD	2002	XI19	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR:England	1916	YEOMAN	G	N-A	G	non-Yr7	G	non-YrSP
2005-2018-UK_RL	GBR	2000	ZEBEDEE	G	N-A	G	non-Yr7	N-A	non-YrSP
2005-2018-UK_RL	GBR	2014	ZULU	G	N-A	G	non-Yr7	G	non-YrSP

				Yr7				*YrSP*	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	*Yr7*	*Yr5 /YrSP* KASP	*YrSP*
core-Watkins	India		W034	C	G	HET	*Yr7*	N-T	-
core-Watkins	India		W126	C	G	HET	*Yr7*	N-T	-
core-Watkins	Greece		W281	C	G	A	*Yr7*	N-T	-
core-Watkins	Spain		W387	C	G	A	*Yr7*	N-T	-
core-Watkins	Portugal		W397	C	G	A	*Yr7*	N-T	-
core-Watkins	India		W433	C	G	HET	*Yr7*	N-T	-
core-Watkins	Afghanistan		W460	C	G	HET	*Yr7*	N-T	-
core-Watkins	Morocco		W496	C	G	A	*Yr7*	N-T	-
core-Watkins	India		W694	C	G	HET	*Yr7*	N-T	-
core-Watkins	Iran		W729	C	G	A	*Yr7*	N-T	-
core-Watkins	India		W732	C	G	HET	*Yr7*	N-T	-
core-Watkins	Iraq		W004	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Australia		W007	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Australia		W023	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W032	G	A	G	non-*Yr7*	N-T	-
core-Watkins	France		W040	G	A	G	non-*Yr7*	N-T	-
core-Watkins	France		W042	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Morocco		W044	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Syria		W045	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W079	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W081	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W092	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Italy		W103	G	A	G	non-*Yr7*	N-T	-
core-Watkins	France		W110	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W127	G	A	G	non-*Yr7*	N-T	-
core-Watkins	France		W139	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W141	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W145	G	A	G	non-*Yr7*	N-T	-
core-Watkins	United Kingdom		W149	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W160	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Poland		W181	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W199	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Egypt		W209	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Morocco		W216	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Tunisia		W218	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W219	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Burma		W223	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W224	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Hungary		W231	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W239	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W246	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Morocco		W254	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Canary Islands		W264	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W273	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Cyprus		W291	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Cyprus		W292	G	HET	G	non-*Yr7*	N-T	-
core-Watkins	Turkey		W299	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Turkey		W300	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Egypt		W305	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Iran		W308	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W315	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W324	G	A	G	non-*Yr7*	N-T	-
core-Watkins	United Kingdom		W325	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Bulgaria		W349	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Yugoslavia		W352	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Yugoslavia		W355	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Yugoslavia		W360	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Portugal		W396	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Palestine		W398	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W406	G	A	G	non-*Yr7*	N-T	-
core-Watkins	India		W420	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W440	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W444	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Romania		W451	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Afghanistan		W468	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Afghanistan		W471	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Afghanistan		W474	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Afghanistan		W475	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Poland		W481	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Poland		W483	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Australia		W507	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W546	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W551	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Greece		W560	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Greece		W562	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Greece		W566	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W568	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Iran		W579	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Iran		W580	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Portugal		W591	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Greece		W605	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Bulgaria		W624	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Iran		W627	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Iran		W629	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Turkey		W637	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Crete		W639	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W651	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W652	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Romania		W662	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Poland		W670	G	A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W671	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Italy		W680	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W683	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Spain		W685	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Greece		W690	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W698	G	A	G	non-*Yr7*	N-T	-
core-Watkins	China		W700	G	A	G	non-*Yr7*	N-T	-
core-Watkins	Iran		W704	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Iran		W705	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	India		W707	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	China		W722	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	India		W731	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	#N/A		W736	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W740	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Algeria		W742	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W746	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Ethiopia		W747	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W749	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W750	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W753	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W771	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Finland		W777	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Italy		W784	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W788	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	USSR		W789	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Tunisia		W811	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Tunisia		W814	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Italy		W816	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	China		W827	G	N-A	G	non-*Yr7*	N-T	-
core-Watkins	Hungary		W912	G	N-A	G	non-*Yr7*	N-T	-

				Yr7				*YrSP*	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	*Yr7*	*Yr5 /YrSP* KASP	*YrSP*
Gediflux	SWE	1945	Virtus	C	G	A	*Yr7*	N-T	-
Gediflux	BEL	1957	Prima	C	G	A	*Yr7*	N-T	-
Gediflux	NLD	1964	Flevina	C	G	A	*Yr7*	N-T	-
Gediflux	FRA	1969	Bouquet (Bouqet	C	G	A	*Yr7*	N-T	-
Gediflux	FRA	1969	Hardi	C	G	A	*Yr7*	N-T	-
Gediflux	NLD	1970	Lely	C	G	A	*Yr7*	N-T	-
Gediflux	FRA	1973	Talent	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1976	Sportsman	C	G	A	*Yr7*	N-T	-
Gediflux	FRA	1980	Camp Rémy	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1982	sabre	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1983	depot	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1983	renard	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1985	Brock	C	G	A	*Yr7*	N-T	-
Gediflux	FRA	1985	soleil	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1987	tara	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1991	Spark	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1992	prophet	C	G	A	*Yr7*	N-T	-
Gediflux			ritz	C	G	A	*Yr7*	N-T	-
Gediflux			trafalgar	C	G	A	*Yr7*	N-T	-
Gediflux	GBR	1978	Mardler	G	G	G	non-*Yr7*	N-T	-
Gediflux	DEU	1990	Toronto	G	G	G	non-*Yr7*	N-T	-
Gediflux	DEU, DNK	1963, 1990	hanno	G	G	G	non-*Yr7*	N-T	-
Gediflux	NLD	1921	Juliana	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1923	Peragis	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1925	Jarl	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1928	Vilmorin 27	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1930	Kadolzer	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1930	Salzmunder Star	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1930	Staring	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1930	Tassilo (Tassilio)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1932	Lovink (Lovenk)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1936	Bersee	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1936	Ebersbacher Wei	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1936	Holdfast	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1937	Criewener 192	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1939	Redman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1939	Rimpaus Bastard	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1939	Rimpaus Braun (	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1940	Carstens 6	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1942	Eroica	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEUDEU	1943	Strubes Dickkopf	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1945	Walde	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1946	Bledor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1946	Cappelle-Desprez	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1946	Hybrid 46	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	CHE:Zurich	1946	Probus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1947	Minister	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1948	Mahndorfer Tem	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1949	baron	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1949	Flanders	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1949	Odin	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1949	Roi Albert	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1949	Schreibers Sturm	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1950	Albatross	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1950	Dr. Lassers Dickk	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1950	Heine 7	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1950	Werla	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1951	Eroica II	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1951	Stamm 101	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1952	Carstens 8	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1952	Steadfast	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1952	Vilmorin 53	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1952	Warden	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1953	Banco	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1953	Ertus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1953	Fanal	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1955	Drauhofener Kolb	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1955	Hochland	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1955	Marco	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1955	Ritzlhofer Neu	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1955	Skandia IIIB (kno	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1956	Svale	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1957	Elite Lepeuple	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1957	Professeur Marc	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1958	Apollo (NL)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1958	Eros	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1958	record	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1959	Admonter	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1959	Champlein	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1959	Loosdorfer Austr	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1959	Starke	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1960	Florian	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1960	Kormoran	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1960	Triumph (A)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1960	Triumph (NL)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1960	Tschermaks Wei	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1961	Cleo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1961	Erla Kolben	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1961	Felix	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1961	Thor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1962	Moisson	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1962	Muck	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1962	Rabe	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	ITA	1962	Regina	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1963	Probstdorfer Stab	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1964	Capitole	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1964	Manella	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1964	Maris Widgeon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1964	Rémois	G	N-A	G	non-*Yr7*	N-T	-

				Yr7				*YrSP*	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	*Yr7*	*Yr5* /*YrSP* KASP	*YrSP*
Gediflux	BEL	1965	Norda	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1965	Rufus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1966	Diplomat	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1966	Joss Cambier	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1966	Multiweiss	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1966	Poros	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1966	Tadorna	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1966	Tambor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1967	Cama	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1967	Cama	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	USA:Montana	1967	crest	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1967	Extrem	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1967	Flinor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1967	Mina	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1968	Caribo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1968	Maris Ranger	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1968	meteor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1968	Starke II (known	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1968	Virgo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1969	Kranich	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1969	sirius	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1970	Fakir	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1970	Orlando	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1970	Top	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1971	Atou	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1971	Maris Huntsman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1971	Maris Nimrod	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1971	Solid	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1972	Aquila	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1972	banner	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1972	Holme	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1972	Mega	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1972	Regent	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1972	Winnetou	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1973	Almus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1973	Benno	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1974	Alcedo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1974	Courtôt	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1974	Danubius	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1974	Lutin	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1974	Maris Freeman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1975	Disponent	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1975	Gamin	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1975	Vuka	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1976	Adam	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1976	Arminda	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1976	Hildur	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1976	Kinsman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1976	wizard	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1977	Hobbit	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1977	Nautica	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1977	Oenus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1977	Pony	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1977	Zemon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1978	Armada (known	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1978	Beauchamp	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	MEX	1978	cheetah	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1978	Fenman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1978	Fidel	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1978	Granta	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1978	Hustler	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1978	Okapi	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1979	Brigand	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1979	Celesta	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1979	David	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1979	Kador	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1979	Virtue	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1979	William	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1980	apostle	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1980	Avalon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1980	breval	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1980	erland	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1980	Granada	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1980	Helge	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1980	Iena	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1980	Jason	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1980	Kanzler	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DNK	1981	Anja	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1981	Compal	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1981	Folke	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1981	guardian	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1981	heinrich	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1981	kronjuwel	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1981	Norman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1981	Pontus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1981	Rapier	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1981	Rektor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1981	Scipion	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	USA:Indiana	1981	Stella	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	USA:Indiana	1981	Stella	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1981	Stetson	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1981	Urban	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1982	Capitaine	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1982	Mission	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1982	Sperber	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1982	Taras	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1982	Titus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1983	Calif	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1983	Galahad	G	N-A	G	non-*Yr7*	N-T	-

				Yr7				*YrSP*	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	*Yr7*	*Yr5 /YrSP* KASP	*YrSP*
Gediflux	AUT	1983	Ikarus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1983	Pernel	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1983	Thésée	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1984	feuvert	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1984	Kosack	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1984	Mercia	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1984	Miras	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL	1984	Odeon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	bert	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	Brimstone	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1985	carolus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1985	corinthian	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1985	Florida	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	Gawain	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	motto	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR,FRA	1985	Moulin	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1985	Obelisk	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	peacock	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	poet	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	prospect	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	rebel	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	rendezvous	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	sickle	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	squadron	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	USA:Colorado	1985	stallion	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1985	vocal	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1985	voyage	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1986	belplaine	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	civic	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	coxswain	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1986	druid	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1986	Expert	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	gambit	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	governor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	Hornet	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1986	Palur	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	patience	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1986	Récital	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	rooster	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1986	Slejpner	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	sniper	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986	trader	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1987	ambassador	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1987	Borenos	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1987	boxer	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1987	dorby	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1987	Escorial	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1987	Faktor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1987	fortress	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1987	parade	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1987	Riband	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1987	sarsen	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1987	Soissons	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1988	fresco	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1988	Haven	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1988	lancelot	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1988	legend	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1988	Mikon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1988	norseman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1988	Orestis	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1988	Pastiche	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1989	axial	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1989	Beaver	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1989	Capo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1989	club	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1989	dean	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1989	Festival	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1989	focus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1989	Greif	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1989	Hereward	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1989	mandate	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1989	Pepital	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1989	Renan	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1989	talon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1989	Zentos	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1990	Contra	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	diablo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1990	estica	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	foreman	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1990	Kontrast	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	leo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	ostara	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	puma	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1990	Ritmo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1990	Sidéral	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	sitka	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	veritas	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1990	wasp	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1991	fletum	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1991	Hunter	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1991	Ibis (D)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1991	Ibis (NL)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1991	orqual	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1991	turpin	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1991	vivant	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1991	zodiac	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	admiral	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	adroit	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	andante	G	N-A	G	non-*Yr7*	N-T	-

				Yr7				*YrSP*	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	*Yr7*	*Yr5* / *YrSP* KASP	*YrSP*
Gediflux	GBR	1992	anthem	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	aristocrat	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1992	athlet	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	buster	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	Cadenza	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	caxton	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	Charger	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	clove	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	fenda	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1992	Genesis	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1992	Georg	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	newhaven	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1992	sarek	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	shannon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	spry	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1992	texel	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	Torfrida	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1992	Trémie	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	welton	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1992	woodstock	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	bandit	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	beaufort	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1993	bercy	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1993	bourbon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	consort	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	dynamo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	encore	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	flash	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1993	fromendor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	hudson	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1993	Lindos	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1993	Meridien	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1993	piccadilly	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	rialto	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	samson	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1993	sennet	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1993	thunder	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1994	alert	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1994	Aztec	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1994	corsaire	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1994	russet	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1994	shango	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1994	victo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1995	Altria	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1995	catamaran	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1995	Equinox	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1995	Flame	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1995	holster	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1995	madrigal	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1995	Optimus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1995	rubens	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1995	Silvius	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE	1995	Stava	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1995	tilburi	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1996	atoll	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR, ITL	1996	chianti	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1996	Flair	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1996	magellan	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1996	Pegassos	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1996	raleigh	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	Abbot	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	asset	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	brutus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	bryden	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	bullet	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1997	Cézanne	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	commodore	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	drake	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	galatea	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1997	Isengrain	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	malacca	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1997	pistol	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	USA:South-Dakota	1997	tandem	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1998	Buchan	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1998	harrier	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	1998	imola	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1998	Savannah	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	1998	semper	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	RUS	1999	Alba	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1999	biscay	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1999	claire	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1999	Napier	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1999	trend	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	2000	deben	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	2001	option	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	2002	access	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	2002	solstice	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	2002	tanker	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD	2002	XI 19	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA	2004	Mendel (known a	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	RUS:Rostov	2006	Terra	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUS:NSW	2011	spitfire	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU		Agron	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			alcier	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU		Apollo (D)	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			blitz	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	CHN		Chinese Spring	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			contour	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			creweau	G	N-A	G	non-*Yr7*	N-T	-

				Yr7				*YrSP*	
				C	G	A		C	
Collection	Origin	Year	Sample	Yr7-A	Yr7-B	Yr7-D	*Yr7*	*Yr5 /YrSP* KASP	*YrSP*
Gediflux	GBR		daphne	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		emblem	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Flair Eigeno Nach	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Glicevka	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Heine 4	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Hesbinion	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Hubertusweizen	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Marisa	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Mironowskaja 80	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Mironowskaja Ju	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Mutant Odeon I	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			Mutant Odeon II	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		rocket	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		saxon	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	SWE		Svalov Kronen	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		torch	G	N-A	G	non-*Yr7*	N-T	-
Gediflux			toucan	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA		Vague d'épis	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU		captor	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		gondola	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		rhino	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA		tessa	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD		tjalk	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		tomo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	AUT	1952, 1986	Hubertus	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1960, 1990	renown	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU	1963, 1990	Bussard	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	DEU, FRA	1970, 1994	Cyrano	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1980, 1992	lynx	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR	1986, 1993	warrior	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		ability	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	BEL		Clovis	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		cobalt	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		destroyer	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		estorial	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	UKR		Flamingo	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	NLD		frista	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		galliard	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA		kontiki	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	FRA		kyalami	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		morell	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		profi	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		rifle	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		spice	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		trawler	G	N-A	G	non-*Yr7*	N-T	-
Gediflux	GBR		wykeham	G	N-A	G	non-*Yr7*	N-T	-

Appendix 8-10. Summary of NLR loci identified in the 10 sequenced wheat genomes (http://www.10wheatgenomes.com).
We first identified NLR loci with NLR-Annotator[195] and then determine whether they were overlapping with a transferred RefSeqv1.1 gene model on a given wheat genome (Gene projections done by http://pgsb.helmholtz-muenchen.de). Where no gene model overlapped with an NLR locus, we used EMBOSS Transeq (https://www.ebi.ac.uk/Tools/st/emboss_transeq/) to generate a 6-frames translation of the NLR loci. We used hmmscan[253] to look for sequence similarity between the NLR protein sequences and conserved domains from the Pfam database (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release). Sequences highlighted in light red were deleted from the set because NLR-Annotator wrongly identified 4 to 5 NLR loci instead of only one. This occurs when there are 'Ns' in the predicted locus. We predicted NLS in sequences highlighted in green with the webtool NLSdb[267] (https://rostlab.org/services/nlsdb/).

NLR-Annotator locus	Overlapping gene model	Alternative	Conserved domains	Comment	Predicted NLS
Arina_1	TraesARI2B01G527000.1		NB		
Arina_2	TraesARI2B01G527100.1		NB		
Arina_3	Arina_3_3	6-frame translation	NB		
Arina_4	TraesARI2B01G527700.1		NB		
Arina_5	Arina_5_3	6-frame translation	NB		
Arina_6	TraesARI2B01G527900.1		NB-LRR		
Arina_7	TraesARI2B01G528100.1		NB-LRR		
Arina_8	Arina_8_3	6-frame translation	NB		
Arina_9	TraesARI2B01G529800.1		BED-NB		Y
Arina_10	TraesARI2B01G530300.1		BED-NB-LRR		Y
Arina_11	Arina_11_1	6-frame translation	BED-NB		
Arina_12				Mis-annotation NLR Annotator	
Arina_13				Mis-annotation NLR Annotator	
Arina_14				Mis-annotation NLR Annotator	
Arina_15	Arina_15_3	6-frame translation	BED-NB		
Arina_16	Arina_16_3	6-frame translation	BED-NB		
Arina_17	TraesARI2B01G531400.1		BED-NB		Y
Jagger_nlr1	TraesJAG2B01G522800.1		NB		
Jagger_nlr2	TraesJAG2B01G522900.1		NB		
Jagger_nlr3	Jagger_3_3	6-frame translation	NB		
Jagger_nlr4	TraesJAG2B01G523500.1		NB		
Jagger_nlr5	Jagger_5_3	6-frame translation	NB		
Jagger_nlr6	TraesJAG2B01G523700.1		NB-LRR		
Jagger_nlr7	TraesJAG2B01G523800.1		NB-LRR		
Jagger_nlr8	TraesJAG2B01G525000.1		NB		
Jagger_nlr9	TraesJAG2B01G525300.1		BED-NB		Y
Jagger_nlr10	TraesJAG2B01G525800.1		BED-NB-LRR		Y
Jagger_nlr11	Jagger_11_2	6-frame translation	BED-NB		
Jagger_nlr12				Mis-annotation NLR Annotator	
Jagger_nlr13				Mis-annotation NLR Annotator	
Jagger_nlr14				Mis-annotation NLR Annotator	
Jagger_nlr15				Mis-annotation NLR Annotator	
Jagger_nlr16	Jagger_16_3	6-frame translation	BED-NB-LRR		
Jagger_nlr17	Jagger_17_3	6-frame translation	BED-NB-LRR		
Jagger_nlr18	TraesJAG2B01G526900.1		BED-NB		Y

NLR-Annotator locus	Overlapping gene model	Alternative	Conserved domains	Comment	Predicted NLS
Julius_1	TraesJUL2B01G522600.1		NB		
Julius_2	TraesJUL2B01G522800.1		NB-LRR		
Julius_3	Julius_3_2	6-frame translation	NB-LRR		
Julius_4	TraesJUL2B01G523700.1 but not taken		BED-NB-LRR	corrected based on Yr5 sequence	Y
Julius_5	Julius_5_3	6-frame translation	BED-NB		
Julius_6	Julius_6_1 and Julius_6_2	6-frame translation	BED-NB		
Julius_7	Julius_7_3	6-frame translation	BED-BED-NB-LRR		
Julius_8	Julius_8_3	6-frame translation	BED-NB		
Julius_9	Julius_9_2	6-frame translation	BED-NB		
Julius_10	Julius_10_3	6-frame translation	BED-NB		
Julius_11	TraesJUL2B01G524700.1		NA	short	
Julius_12	TraesJUL2B01G525000.1		BED-NB		Y
Julius_13	Julius_13_4 and Julius_13_6	6-frame translation	BED-NB		
Lancer_1	TraesLAC2B01G496000.1		NB		
Lancer_2	TraesLAC2B01G496100.1		NB-LRR		
Lancer_3	Lancer_3_2	6-frame translation	NB-LRR		
Lancer_4	Lancer_4_2	6-frame translation	BED-NB-LRR	corrected based on Yr5 sequence	
Lancer_5	Lancer_5_3	6-frame translation	BED-NB		
Lancer_6	TraesLAC2B01G497200.1		BED-NB	numerous stop codons	
Lancer_7	Lancer_7_3	6-frame translation	BED-BED-NB-LRR		
Lancer_8	Lancer_8_3	6-frame translation	BED-NB		
Lancer_9	Lancer_9_3	6-frame translation	BED-NB		
Lancer_10	Lancer_10_3	6-frame translation	BED-NB		
Lancer_11	Lancer_11_3	6-frame translation	BED-NB		
Lancer_12	Lancer_12_3	6-frame translation	BED-NB		
Lancer_13	TraesLAC2B01G498200.1		BED-NB		Y
Landmark_1	TraesLDM2B01G518300		NB		
Landmark_2	TraesLDM2B01G518400		NB		
Landmark_3	Landmark_3_3	6-frame translation	NB		
Landmark_4	TraesLDM2B01G518700		NB		Y
Landmark_5	TraesLDM2B01G519400		NB		
Landmark_6	TraesLDM2B01G519700		BED-NB	numerous stop codons	
Landmark_7	Landmark_7_2 and 7_3	6-frame translation	BED-NB		
Landmark_8	Landmark_8_2	6-frame translation	BED-NB		
Landmark_9	Landmark_9_2 and Landmark_9_3	6-frame translation	BED-NB		
Landmark_10	Landmark_10_1 and Landmark_10_2	6-frame translation	BED-NB	Yr7 alleles so annotated based on Yr7	
Landmark_11	Landmark_11_3	6-frame translation	NB		
Landmark_12	Landmark_12_3	6-frame translation	BED-NB		
Landmark_13	TraesLDM2B01G520400.1		NB-LRR	numerous stop codons	
Landmark_14	Landmark_14_1	6-frame translation	NB		

NLR-Annotator locus	Overlapping gene model	Alternative	Conserved domains	Comment	Predicted NLS
Landmark_15	Landmark_15_1	6-frame translation	BED-NB		
Landmark_16	TraesLDM2B01G520800		BED-NB-LRR	numerous stop codons	
Landmark_17	Landmark_17_3	6-frame translation	NB-NB		
Landmark_18	Landmark_18_1 and 18_3	6-frame translation	BED-NB		
Mace_1	TraesMAC2B01G527300.1		NB		
Mace_2	TraesMAC2B01G527400.1		NB		
Mace_3	Mace_2_3	6-frame translation	NB		
Mace_4	TraesMAC2B01G527900.1		NB		
Mace_5	Mace_5_2	6-frame translation	NB-LRR		
Mace_6	TraesMAC2B01G528900.1		BED-NB	numerous stop codons	
Mace_7	Mace_7_2 and Mace_7_3	6-frame translation	BED-NB-LRR		
Mace_8	Mace_8_2	6-frame translation	BED-NB		
Mace_9	Mace_9_2 and Mace_9_3	6-frame translation	BED-NB		
Mace_10	Mace_10_1 and Mace_10_2	6-frame translation	BED-NB-LRR	Yr7 alleles so annotated based on Yr7	
Mace_11	Mace_11_3	6-frame translation	NB		
Mace_12	Mace_12_3	6-frame translation	BED-NB		
Mace_13	Mace_13_3	6-frame translation	BED-NB-LRR		
Mace_14	Mace_14_3	6-frame translation	NB		
Mace_15	TraesMAC2B01G529600.1		BED-NB	numerous stop codons	
Mace_16	Mace_16_3	6-frame translation	BED-NB-LRR		
Mace_17	Mace_17_3	6-frame translation	NB		
Mace_18	Mace_18_1 and Mace_18_3	6-frame translation	BED-NB		
Norin61_1	TraesNOR2B01G529900.1		NB		
Norin61_2	TraesNOR2B01G530000.1		NB		
Norin61_3	Norin61_3_3	6-frame translation	NB		
Norin61_4	TraesNOR2B01G530200.1		NB		
Norin61_5	Norin61_5_3	6-frame translation	NB		
Norin61_6	TraesNOR2B01G530400.1		NB-LRR		
Norin61_7	TraesNOR2B01G530500.1		NB-LRR		
Norin61_8	Norin61_8_3	6-frame translation	NB		
Norin61_9	TraesNOR2B01G531900.1		BED-NB		Y
Norin61_10	TraesNOR2B01G532400.1		BED-NB-LRR		Y
Norin61_11	Norin61_11_2	6-frame translation	BED-NB		
Norin61_12				Mis-annotation NLR Annotator	
Norin61_13				Mis-annotation NLR Annotator	
Norin61_14				Mis-annotation NLR Annotator	
Norin61_15	Norin61_15_3	6-frame translation	BED-NB-LRR		
Norin61_16	Norin61_16_3	6-frame translation	BED-NB-LRR		
Norin61_17	TraesNOR2B01G533500.1		BED-NB		Y
Stanley_1	TraesSTA2B01G531500.1		NB		
Stanley_2	TraesSTA2B01G531600.1		NB		
Stanley_3	Stanley_3_3	6-frame translation	NB		
Stanley_4	TraesSTA2B01G532100.1		NB		

NLR-Annotator locus	Overlapping gene model	Alternative	Conserved domains	Comment	Predicted NLS
Stanley_5	Stanley_5_2	6-frame translation	NB-LRR		
Stanley_6	TraesSTA2B01G533100.1		BED-NB	numerous stop codons	
Stanley_7	Stanley_7_1 and Stanley_7_3	6-frame translation	BED-NB-LRR		
Stanley_8	Stanley_8_2	6-frame translation	BED-NB		
Stanley_9	Stanley_9_2 and Stanley_9_3	6-frame translation	BED-NB		
Stanley_10	Stanley_10_1 and Stanley_10_2	6-frame translation	BED-NB-LRR	Yr7 alleles so annotated based on Yr7	
Stanley_11	Stanley_11_1	6-frame translation	BED-NB		
Stanley_12	Stanley_12_3	6-frame translation	BED-NB-LRR		
Stanley_13	Stanley_13_3	6-frame translation	NB		
Stanley_14	Stanley_14_1 and Stanley_14_3	6-frame translation	BED-NB		
Stanley_15	Stanley_15_6	6-frame translation	NB-LRR		
Stanley_16	Stanley_16_4	6-frame translation	NB		
Stanley_17	Stanley_17_5	6-frame translation	BED-NB		
Stanley_18	TraesSTA2B01G533400.1		BED*-NB-LRR	numerous stop codons	
SY_Mattis_1	TraesSYM2B01G522400.1		NB		
SY_Mattis_2	TraesSYM2B01G522500.1		NB		
SY_Mattis_3	SY_Mattis_3_3	6-frame translation	NB		
SY_Mattis_4	TraesSYM2B01G523000.1		NB		
SY_Mattis_5	SY_Mattis_5_2	6-frame translation	NB-LRR		
SY_Mattis_6	SY_Mattis_6_1 and SY_Mattis_6_2	6-frame translation	BED-NB	numerous additional domains in SY_Mattis_6_2	
SY_Mattis_7	SY_Mattis_7_1 and SY_Mattis_7_3	6-frame translation	BED-NB-LRR		
SY_Mattis_8	SY_Mattis_8_2	6-frame translation	BED-NB		
SY_Mattis_9	SY_Mattis_9_2 and SY_Mattis_9_33	6-frame translation	BED-NB		
SY_Mattis_11	SY_Mattis_11_3	6-frame translation	BED-NB		
SY_Mattis_12	TraesSYM2B01G524900.1		NB-LRR	numerous stop codons	
SY_Mattis_13	SY_Mattis_13_3	6-frame translation	NB		
SY_Mattis_14	TraesSYM2B01G525000.1		BED-NB	numerous stop codons	
SY_Mattis_15	SY_Mattis_15_3	6-frame translation	BED-NB-LRR		
SY_Mattis_17	SY_Mattis_17_1 and SY_Mattis_17_2	6-frame translation	BED-NB		
CS_nlr1	TraesCS2B02G486100.1		NB		
CS_nlr2	TraesCS2B02G486200.1		NB		
CS_nlr3	TraesCS2B02G486247.1		NA	short	
CS_nlr4	TraesCS2B02G486300.1		NB		
CS_nlr5	TraesCS2B02G486390.1		NA	short	
CS_nlr6	TraesCS2B02G486400.1		NB		
CS_nlr7	TraesCS2B02G486700.1		NB		
CS_nlr8	CS_8_3	6-frame translation	NB		
CS_nlr9	TraesCS2B02G488000.1		BED-NB		Y
CS_nlr10	TraesCS2B02G488400.1		BED-NB		Y
CS_nlr11	CS_11_2	6-frame translation	BED-NB		
CS_nlr15	CS_15_3	6-frame translation	BED-NB-LRR		

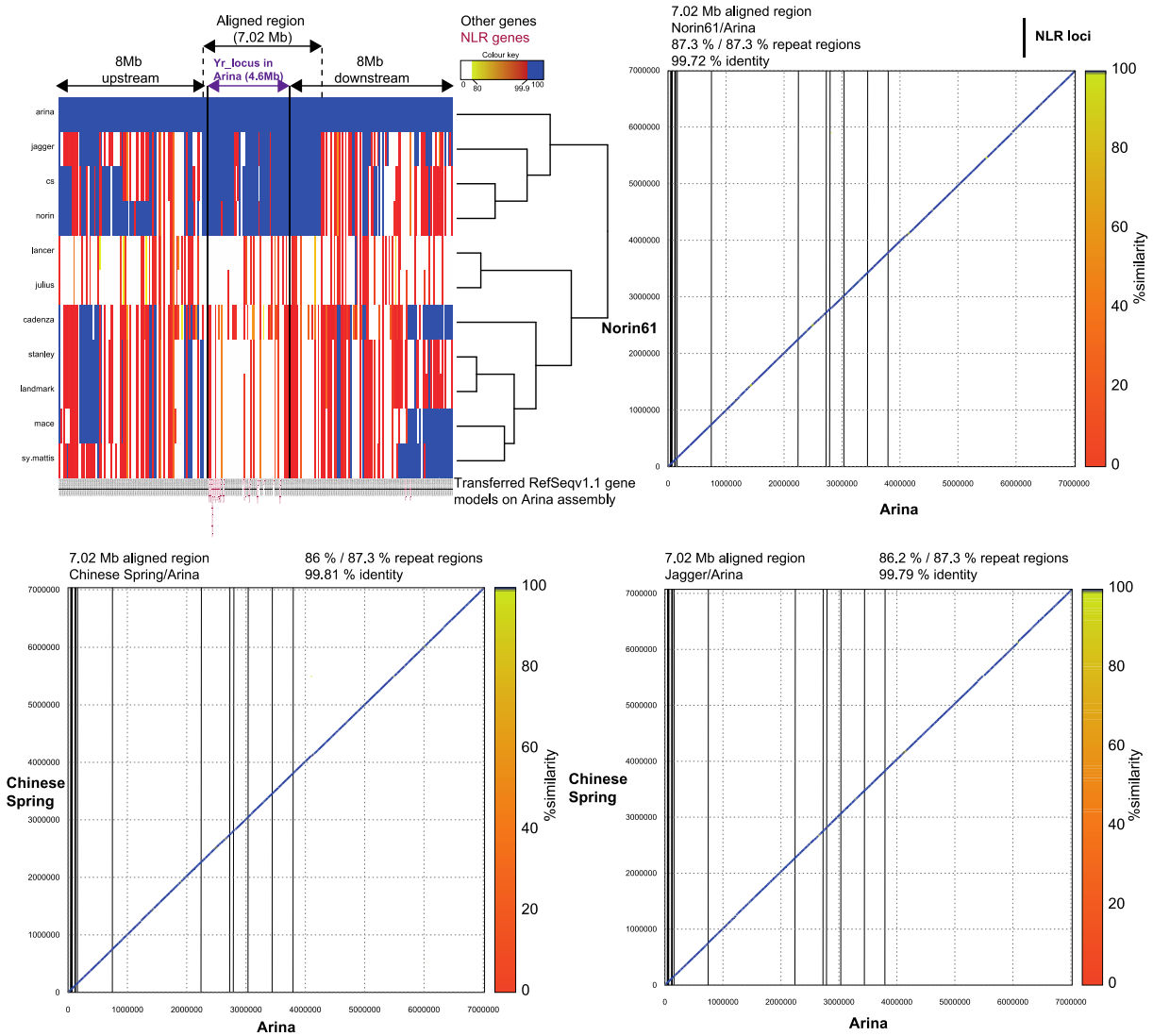
NLR-Annotator locus	Overlapping gene model	Alternative	Conserved domains	Comment	Predicted NLS
CS_nlr16	CS_16_3	6-frame translation	BED-NB-LRR		
CS_nlr17	TraesCS2B02G489400.1		BED-NB		Y
Yr7	annotated in Marchal et al., 2018		BED-NB-LRR		
Yr5	annotated in Marchal et al., 2018		BED-NB-LRR		
YrSP	annotated in Marchal et al., 2018		BED-NB-LRR		

Appendix 8-11. Alignment statistics derived from MUMmer (v3.0) analysis in the expanded *Yr* region

	Length of selected reference (Mb)	Length of aligned reference (Mb)	Length of selected query (Mb)	Length of aligned (Mb)
Julius/Lancer	12.85	12.64	12.9	12.71
Landmark/SY-mattis	9.7	9.5	9.6	9.4
Landmark/Mace	9.7	9	11.1	9.6
Landmark/Stanley	9.7	9.5	9.7	9.5
Arina/Chinese-Spring	7.02	6.95	7.04	6.92
Arina/Jagger	7.02	6.95	7.07	6.96
Arina/Norin61	7.02	6.97	7.02	6.96

	Length of *Yr* locus in reference	%ID (upstream *Yr* locus)	%ID (*Yr* locus)	%ID (downstream *Yr* locus)
Julius/Lancer	2.2	99.959	99.902	99.954
Landmark/SY-mattis	3.6	99.643	99.925	99.948
Landmark/Mace	3.6	95.515	99.972	99.977
Landmark/Stanley	3.6	99.609	99.329	99.958
Arina/Chinese-Spring	4.6	99.947	99.807	99.813
Arina/Jagger	4.6	98.902	99.814	99.788
Arina/Norin61	4.6	99.779	99.684	99.934

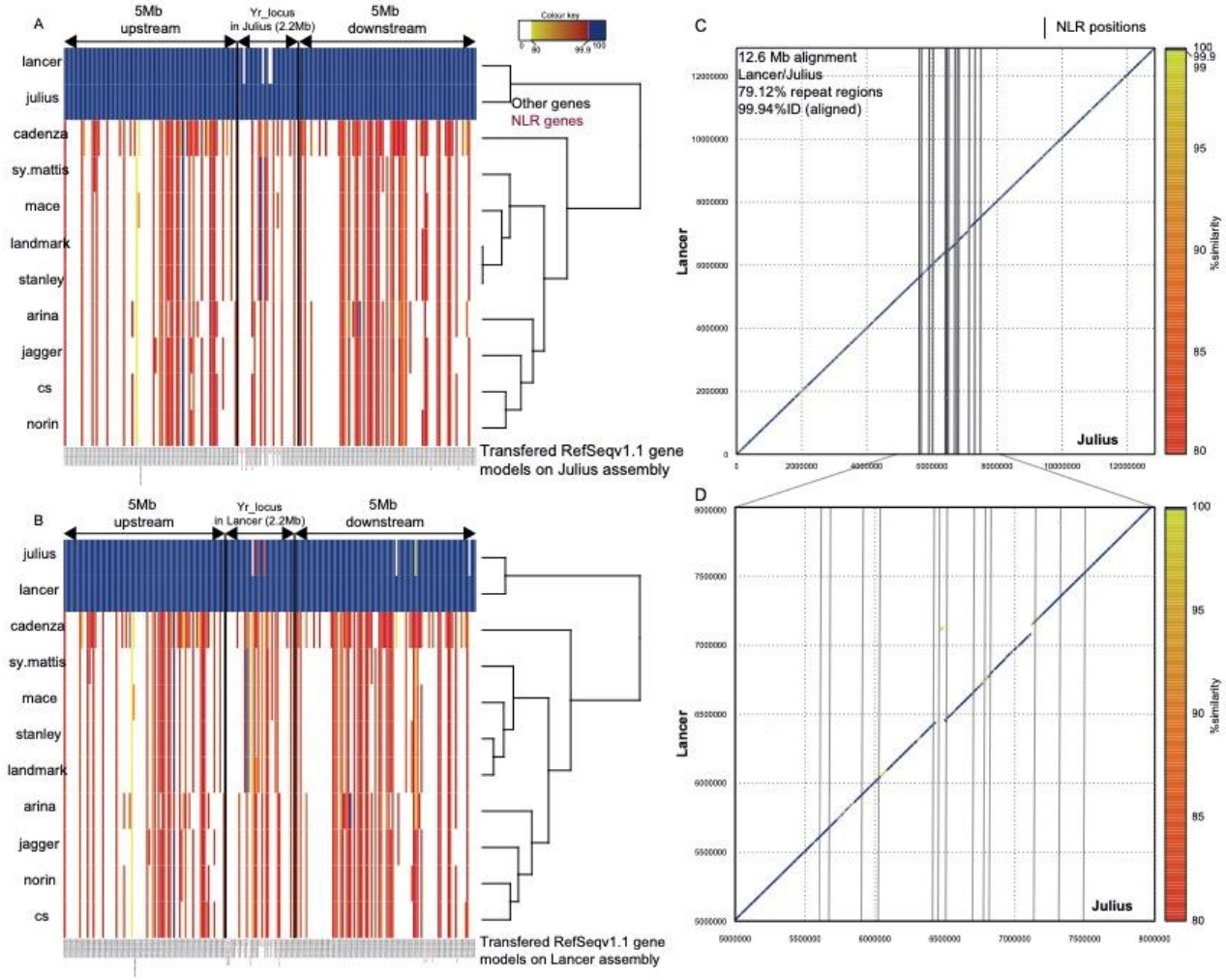
	#SNPs incl InDel (excluding Ns)	SNP density (#SNP/Mb)	#SNPs incl InDel (upstream *Yr* locus)	#SNPs incl InDel (*Yr* locus)	SNP density *Yr* locus (#SNP/Mb)	Weighted #SNPs based on *Yr* locus SNP density (difference with observed #SNP)	#SNPs incl InDel (downstream Yr locus)
Julius/Lancer	266	21	68	137	62	784 (+518)	61
Landmark/SY-mattis	745	78	372	172	48	456 (-289)	201
Landmark/Mace	3651	406	3595	19	5	45 (-3606)	37
Landmark/Stanley	348	37	283	26	7	67 (-281)	39
Arina/Chinese-Spring	2639	380	915	1712	372	2586 (-53)	12
Arina/Jagger	2760	397	949	1799	391	2718 (-42)	12
Arina/Norin61	2883	414	967	1904	414	2886 (+ 3)	12



Appendix 8-12. BLAST analysis between Arina and the nine other wheat genomes + Cadenza (heatmap, top left) and alignments of the corresponding region (dashed line) in Norin61 (top right), Chinese Spring (bottom left) and Jagger (bottom right).

Heatmap: Only hits that overlapped at least 90 % of the query and located on Chromosome 2B are displayed. The colour key ranges from white (no hit to < 80 % identity hits), yellow (close to 80 % identity hits), red (close to 99.9 % identity hits) to blue (strictly 100 % identity hits). The black arrows on at the top show the boundaries of the region with the location of the Yr region. The black dashed line shows the boundaries of the aligned region between Arina and Norin61 (right). Gene identifiers are displayed on at the bottom with black indicating a non-NLR locus and dark red a NLR locus. Results were clustered according to the hierarchical clustering methods implemented in the heatmap 2 function of R (gplots package, v3.0.1.1.).

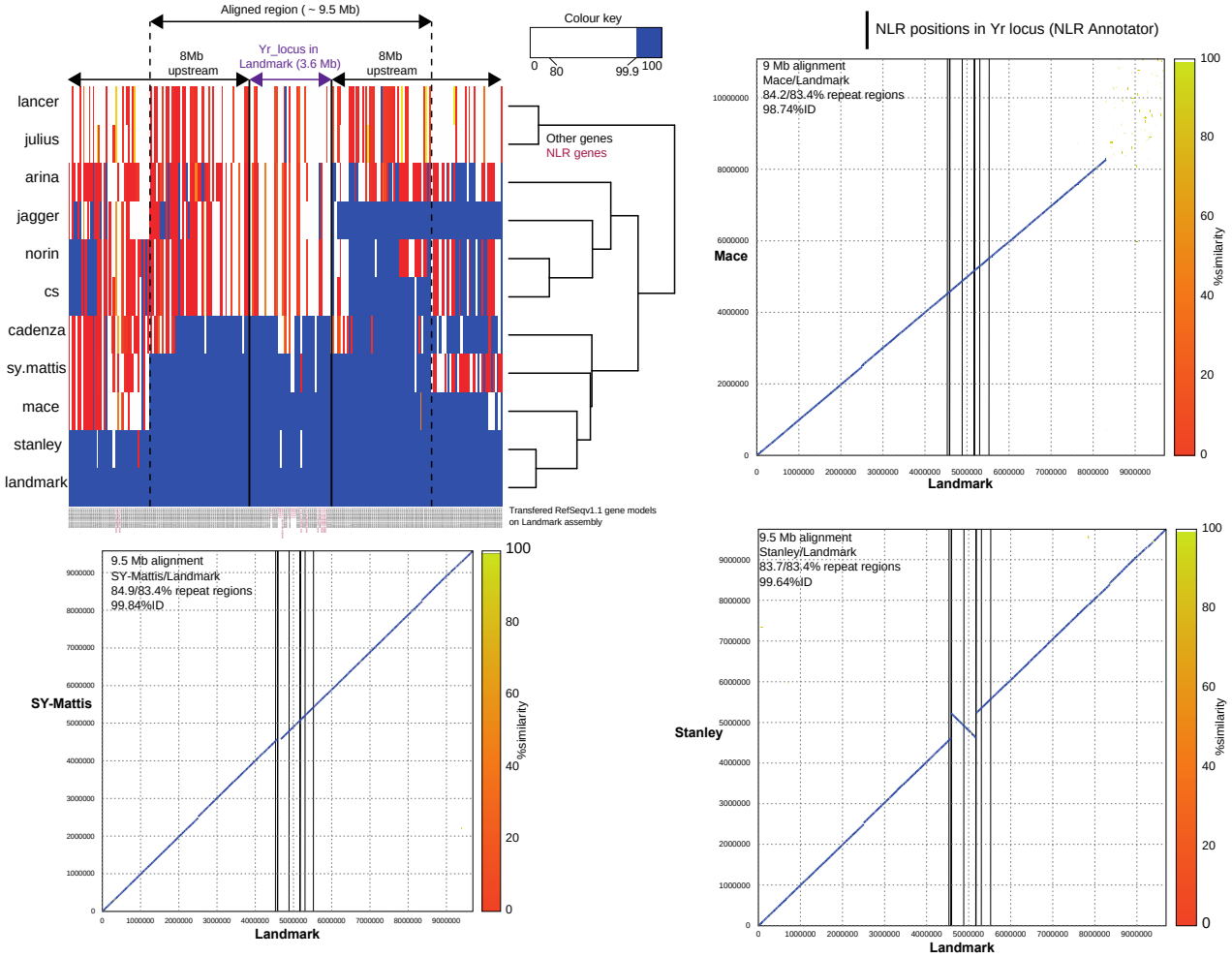
Alignments: alignment of the 7.02 Mb region defined from the heatmap (left) as the most conserved part between Arina, Chinese Spring, Jagger and Norin61. Blue colour shows region sharing a percentage identity higher than 99.9 %. Black vertical lines indicate the position of the NLR loci. The alignment was performed with MUMmer v3.0



Appendix 8-13. BLAST analysis between Julius and the nine other wheat genomes (top left) and between Lancer and the nine other wheat genomes (bottom left). Alignment between Julius and Lancer of the whole region (top right) and corresponding close-up in the *Yr* region (bottom right)

Heatmaps: Only hits that overlapped at least 90 % of the query and located on Chromosome 2B are displayed. The colour key ranges from white (no hit to < 80 % identity hits), yellow (close to 80 % identity hits), red (close to 99.9 % identity hits) to blue (strictly 100 % identity hits). The black arrows on at the top show the boundaries of the region with the location of the *Yr* region. Gene identifiers are displayed on at the bottom with black indicating a non-NLR locus and dark red a NLR locus. Results were clustered according to the hierarchical clustering methods implemented in the heatmap 2 function of R (gplots package, v3.0.1.1.).

Alignments: alignment of the 12.6 Mb region defined from the heatmap (left). Blue colour shows region sharing a percentage identity higher than 99.9 %. Black vertical lines indicate the position of the NLR loci. The alignment was performed with MUMmer v3.0.



Appendix 8-14. BLAST analysis between Landmark and the nine other wheat genomes + Cadenza (heatmap, top left) and alignments of the corresponding region (dashed line) in Mace (top right), SY-Mattis (bottom left) and Stanley (bottom right).

Heatmap: Only hits that overlapped at least 90 % of the query and located on Chromosome 2B are displayed. The colour key ranges from white (no hit to < 80 % identity hits), yellow (close to 80 % identity hits), red (close to 99.9 % identity hits) to blue (strictly 100 % identity hits). The black arrows on at the top show the boundaries of the region with the location of the *Yr* region. The black dashed line shows the boundaries of the aligned region between Landmark and the three other varieties Mace, SY-Mattis and Stanley (right). Gene identifiers are displayed on at the bottom with black indicating a non-NLR locus and dark red a NLR locus. Results were clustered according to the hierarchical clustering methods implemented in the heatmap 2 function of R (gplots package, v3.0.1.1.).

Alignments: alignment of the 9.7 Mb region defined from the heatmap (left) as the most conserved part between Landmark, Mace, SY-Mattis and Stanley. Blue colour shows region sharing a percentage identity higher than 99.9 %. Black vertical lines indicate the position of the NLR loci. The alignment was performed with MUMmer v3.0

Appendix 8-15. Definition of the syntenic region across different grasses (see Table below) and identified NLR loci with NLR Annotator

				Orthologs retrieved from Ensembl			%aligned > 98 && %ID > 95			%aligned > 98 && %ID > 95			%aligned > 98 && %ID > 95	%aligned > 98 && %ID > 95
				%ID > 76	%ID > 84	%ID > 92	Aegilops tauschii			Triticum turgidum Zavitan			RefSeqv1.0 A subgenome	RefSeqv1.0 D subgenome
RefSeqv1.0				Oryza sativa Japonica	Brachypodium Distachyon	Hordeum vulgare	Chromosome	start	end	Chromosome	start	end	Gene ID	Gene ID
chr2B	682190941	682192727	TraesCS2B01G485000											
chr2B	682217665	682219365	TraesCS2B01G485100										TraesCS2A01G622000LC	TraesCS2D01G464200
chr2B	682258533	682260001	TraesCS2B01G485300							chr2B	680745370	680745765		
chr2B	682572588	682573604	TraesCS2B01G485400											
chr2B	682741131	682745018	TraesCS2B01G485500											
chr2B	682848631	682859742	TraesCS2B01G485600											
chr2B	683002069	683004931	TraesCS2B01G485700	OS04G0620200	BRADI5G22090	HORVU2Hr1G103560								
chr2B	683021882	683029367	TraesCS2B01G485800											
chr2B	683035392	683040011	TraesCS2B01G486000				Chr2	567952017	567956640					
chr2B	683043705	683047767	Ta_2B1											
chr2B	683054818	683058958	Ta_2B2											
chr2B	683063722	683064832	Ta_2B3											
chr2B	683068415	683073736	Ta_2B4											
chr2B	683116693	683118802	Ta_2B5											
chr2B	683128929	683133337	Ta_2B6											
chr2B	683160286	683163016	Ta_2B7											
chr2B	683174150	683176245	TraesCS2B01G486900											
chr2B	683483762	683487437	TraesCS2B01G487100											
chr2B	683752036	683756176	Ta_2B8											
chr2B	685146969	685148388	TraesCS2B01G487900											
chr2B	685266071	685270775	Ta_2B9											
chr2B	685502824	685503358	TraesCS2B01G488200											
chr2B	685742944	685746490	Ta_2B10											
chr2B	686047270	686050687	Ta_2B11											
chr2B	686055835	686056522	TraesCS2B01G488800											
chr2B	686455267	686458864	Ta_2B12											
chr2B	686464658	686465562	TraesCS2B01G489100											
chr2B	686811712	686815294	Ta_2B13											
chr2B	686818413	686820009	TraesCS2B01G489500				Chr2	569768288	569769901					
chr2B	686834386	686838510	TraesCS2B01G489600			HORVU2Hr1G103310								
chr2B	687204794	687209908	TraesCS2B01G489700											
chr2B	687469319	687473404	TraesCS2B01G490100	OS04G0623300	BRADI5G22220									
chr2B	687634061	687635470	TraesCS2B01G490200											
chr2B	688060256	688061752	TraesCS2B01G490300											
chr2B	688060256	688061752	TraesCS2B01G490400											
chr2B	688239893	688242651	TraesCS2B01G490500											
chr2B	688375493	688377107	TraesCS2B01G490600											
chr2B	688430739	688433903	TraesCS2B01G490700											
chr2B	688456751	688459118	TraesCS2B01G490800							chr2B	685426399	685426213	TraesCS2A01G467800	TraesCS2D01G467800

Specie	Cultivar/group	Source	Link/ref
Triiticum aestivum	Chinese Spring	IWGSC	https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies
Triticum turgidum	Zavitan	WEWseq	Avni et al. 2017[23]
Aegilops tauschii		UC Davis	Luo et al. 2017[255]
Oryza sativa	Japonica	Ensembl / RAP-DB	http://plants.ensembl.org/Oryza_sativa/Info/Index
Brachypodium distachyon		Ensembl / Brachypodium.org	http://plants.ensembl.org/Brachypodium_distachyon/Info/Index
Hordeum vulgare	Morex	Ensembl / IBSC	http://plants.ensembl.org/Hordeum_vulgare/Info/Index

Appendix 8-16. Identified BED-containing proteins in RefSeq v1.0 based on a hmmer scan analysis (see Methods 4.2.5).
Several features are added: number of identified BED domains and the presence of other conserved domains present, the best BLAST hit from the non-redundant database of NCBI with its description and score, and whether the BED domain was related to BED domains from NLR proteins based on the neighbour network shown in Figure 4-14.

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search hmmer /	CD-Search hmmer /	CD-Search hmmer /	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
TraesCS1A01G002600.2	1	ZnF_BED					EMS48536.1	hypothetical protein TRIUR3_00706 [Triticum urartu]	80	
TraesCS1B01G130400.1	1	ZnF_BED	DUF659	Dimer_Tnp_hAT			XP_023157898.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Zea mays]	48.977	
TraesCS1B01G158800.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_016740977.1	PREDICTED: zinc finger BED domain-containing protein RICESLEEPER 2-like [Gossypium hirsutum]	42.837	Yes
TraesCS1B01G454300.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020157945.1	zinc finger BED domain-containing protein RICESLEEPER 2-like isoform X1 [Aegilops tauschii subsp. tauschii]	59.23	
TraesCS1B01G475300.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020157945.1	zinc finger BED domain-containing protein RICESLEEPER 2-like isoform X1 [Aegilops tauschii subsp. tauschii]	58.748	
TraesCS1D01G004800.1	1	ZnF_BED	DnaJ				XP_020149352.1	uncharacterized protein LOC109734557 isoform X1 [Aegilops tauschii subsp. tauschii]	100	
TraesCS1D01G137700.1	1	ZnF_BED					XP_020167145.1	zinc finger BED domain-containing protein RICESLEEPER 3-like isoform X1 [Aegilops tauschii subsp. tauschii]	73.826	
TraesCS2A01G246500.1	1	ZnF_BED					EMS59629.1	hypothetical protein TRIUR3_24222 [Triticum urartu]	97.38	
TraesCS2A01G477100.1	2	ZnF_BED	ZnF_BED	DUF4413	Dimer_Tnp_hAT		EMS55659.1	Putative AC transposase [Triticum urartu]	99.635	
TraesCS2A01G477800.1	2	ZnF_BED	ZnF_BED	DUF4413	Dimer_Tnp_hAT		XP_020151639.1	zinc finger BED domain-containing protein RICESLEEPER 2-like	89.486	

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search hmmer /	CD-Search hmmer /	CD-Search hmmer /	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
								[Aegilops tauschii subsp. tauschii]		
TraesCS2A01G478200.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020151658.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	89.816	
TraesCS2B01G151300.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_015649039.1	PREDICTED: zinc finger BED domain-containing protein RICESLEEPER 2-like [Oryza sativa Japonica Group]	45.997	
TraesCS2B01G262500.1	1	ZnF_BED					XP_020167145.1	zinc finger BED domain-containing protein RICESLEEPER 3-like isoform X1 [Aegilops tauschii subsp. tauschii]	76.684	
TraesCS2B01G442000.1	1	ZnF_BED	DUF659				XP_020193659.1	uncharacterized protein LOC109779449 [Aegilops tauschii subsp. tauschii]	72.761	
TraesCS2B01G500400.1	2	ZnF_BED	ZnF_BED	DUF4413	Dimer_Tnp_hAT		EMS55659.1	Putative AC transposase [Triticum urartu]	96.727	
TraesCS2B01G501100.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020151639.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	94.286	
TraesCS2B01G509900.1	1	ZnF_BED					CDM80226.1	unnamed protein product [Triticum aestivum]	75.41	
TraesCS2B01G574700.1	1	ZnF_BED	DUF659				GAV86993.1	LOW QUALITY PROTEIN: zf-BED domain-containing protein/DUF659 domain-containing protein/Dimer_Tnp_hAT domain-containing protein	62.419	
TraesCS2D01G476500.1	2	ZnF_BED	ZnF_BED	DUF4413	Dimer_Tnp_hAT		XP_020151643.1	zinc finger BED domain-containing protein RICESLEEPER 1-like [Aegilops tauschii subsp. tauschii]	99.879	
TraesCS2D01G477000.1	2	ZnF_BED	ZnF_BED	DUF4413	Dimer_Tnp_hAT		XP_020151639.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	99.771	

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search hmmer /	CD-Search hmmer /	CD-Search hmmer /	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
TraesCS2D01G477400.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020151658.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	98.951	
TraesCS3A01G018400.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020157945.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	60.16	
TraesCS3A01G214600.1	1	ZnF_BED	PXB superfamily N	Dimer_Tnp_hAT			XP_020394920.1	zinc finger BED domain-containing protein RICESLEEPER 1-like isoform X1 [Zea mays]	45.902	
TraesCS3A01G256500.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020169532.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	98.857	
TraesCS3A01G304000.1	1	ZnF_BED					XP_012704570.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Setaria italica]	55.556	
TraesCS3A01G406300.1	1	ZnF_BED					XP_014755075.1	PREDICTED: uncharacterized protein LOC104583357 isoform X2 [Brachypodium distachyon]	44.203	
TraesCS3A01G459900.1	1	ZnF_BED					KQJ95646.1	hypothetical protein BRADI_3g18330v3 [Brachypodium distachyon]	40.397	
TraesCS3B01G231800.2	2	ZnF_BED	PHD_SF	PHD_SF			XP_020154732.1	uncharacterized protein LOC109740111 isoform X3 [Aegilops tauschii subsp. tauschii]	96.753	
TraesCS3B01G238000.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			CDM80226.1	unnamed protein product [Triticum aestivum]	72.131	
TraesCS3B01G269600.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020177565.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	94.43	yes
TraesCS3B01G289700.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020169532.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	98.571	

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search hmmer /	CD-Search hmmer /	CD-Search hmmer /	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
TraesCS3B01G317800.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020177565.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	92.911	yes
TraesCS3B01G440000.1	1	ZnF_BED					XP_014755075.1	PREDICTED: uncharacterized protein LOC104583357 isoform X2 [Brachypodium distachyon]	41.212	
TraesCS3B01G499600.1	1	ZnF_BED					KQJ95646.1	hypothetical protein BRADI_3g18330v3 [Brachypodium distachyon]	32.353	
TraesCS3D01G026300.2	1	ZnF_BED	DUF4413				XP_020180515.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	95.385	
TraesCS3D01G152700.1	1	ZnF_BED					EMS68829.1	hypothetical protein TRIUR3_25498 [Triticum urartu]	82.353	
TraesCS3D01G202300.1	1	ZnF_BED	PHD_SF	PHD_SF			XP_020154732.1	uncharacterized protein LOC109740111 isoform X3 [Aegilops tauschii subsp. tauschii]	99.721	
TraesCS3D01G256800.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020169532.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	100	
TraesCS3D01G350600.1	1	ZnF_BED	DUF659	Dimer_Tnp_hAT			XP_020167372.1	uncharacterized protein LOC109752869 [Aegilops tauschii subsp. tauschii]	85.241	
TraesCS3D01G401500.1	1	ZnF_BED					XP_014755075.1	PREDICTED: uncharacterized protein LOC104583357 isoform X2 [Brachypodium distachyon]	47.059	
TraesCS3D01G452700.1	1	ZnF_BED					KQJ95646.1	hypothetical protein BRADI_3g18330v3 [Brachypodium distachyon]	33.702	
TraesCS4A01G068800.1	2	ZnF_BED	ZnF_BED	Sin_N superfamily NC	DUF4413	Dimer_Tnp_hAT	XP_020197808.1	zinc finger BED domain-containing protein RICESLEEPER 2-like isoform X1 [Aegilops tauschii subsp. tauschii]	98.25	
TraesCS4A01G069300.1	1	ZnF_BED					XP_012699703.1	zinc finger BED domain-containing protein	52.809	

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search / hmmer	CD-Search / hmmer	CD-Search / hmmer	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
								RICESLEEPER 2-like [Setaria italica]		
TraesCS4A01G251000.1	2	ZnF_BED	ZnF_BED	DUF4413	Dimer_Tnp_hAT		XP_020201069.1	zinc finger BED domain-containing protein RICESLEEPER 2 [Aegilops tauschii subsp. tauschii]	99.319	
TraesCS4A01G330500.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			AAM74247.1	Putative transposable element [Oryza sativa Japonica Group]	62.578	
TraesCS4B01G012600.1	1	ZnF_BED	DUF659				XP_020178553.1	uncharacterized protein LOC109764115 [Aegilops tauschii subsp. tauschii]	84.581	
TraesCS4B01G063700.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020148416.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	98.711	
TraesCS4B01G063900.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020201069.1	zinc finger BED domain-containing protein RICESLEEPER 2 [Aegilops tauschii subsp. tauschii]	98.365	
TraesCS4B01G163200.1	1	ZnF_BED					XP_021971416.1	zinc finger BED domain-containing protein RICESLEEPER 3-like [Helianthus annuus]	45.455	
TraesCS4B01G225900.2	1	ZnF_BED	ZnF_BED	Sin_N superfamily NC	DUF4413	Dimer_Tnp_hAT	XP_020197808.1	zinc finger BED domain-containing protein RICESLEEPER 2-like isoform X1 [Aegilops tauschii subsp. tauschii]	96.504	
TraesCS4B01G264000.1	1	ZnF_BED					PAN37863.1	hypothetical protein PAHAL_G00842 [Panicum hallii]	47.863	
TraesCS4D01G062600.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020148416.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	100	
TraesCS4D01G062700.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020201069.1	zinc finger BED domain-containing protein RICESLEEPER 2 [Aegilops tauschii subsp. tauschii]	99.727	
TraesCS4D01G131800.1	1	ZnF_BED					XP_021971416.1	zinc finger BED domain-containing protein RICESLEEPER 3-like [Helianthus annuus]	45.455	

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search hmmer /	CD-Search hmmer /	CD-Search hmmer /	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
TraesCS4D01G226600.3	1	ZnF_BED	Sin_N superfamily NC	DUF4413	Dimer_Tnp_hAT		XP_020197808.1	zinc finger BED domain-containing protein RICESLEEPER 2-like isoform X1 [Aegilops tauschii subsp. tauschii]	99.875	
TraesCS5A01G144500.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020180515.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	95.421	
TraesCS5A01G200200.1	1	ZnF_BED					XP_020169880.1	zinc finger BED domain-containing protein DAYSLEEPER-like [Aegilops tauschii subsp. tauschii]	92.009	
TraesCS5A01G200400.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020190052.1	zinc finger BED domain-containing protein RICESLEEPER 3-like [Aegilops tauschii subsp. tauschii]	97.329	
TraesCS5A01G273600.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			AAP52341.1	hAT family dimerisation domain containing protein [Oryza sativa Japonica Group]	61.874	
TraesCS5A01G483700.1	1	ZnF_BED	DUF659				XP_020173835.1	uncharacterized protein LOC109759423 isoform X2 [Aegilops tauschii subsp. tauschii]	90.196	
TraesCS5A01G527200.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020157945.1	zinc finger BED domain-containing protein RICESLEEPER 2-like isoform X1 [Aegilops tauschii subsp. tauschii]	60.64	
TraesCS5B01G001200.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_015619043.1	PREDICTED: zinc finger BED domain-containing protein RICESLEEPER 1-like [Oryza sativa Japonica Group]	39.187	
TraesCS5B01G004100.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_015619043.1	PREDICTED: zinc finger BED domain-containing protein RICESLEEPER 1-like [Oryza sativa Japonica Group]	39.187	
TraesCS5B01G198600.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020169880.1	zinc finger BED domain-containing protein DAYSLEEPER-like [Aegilops tauschii subsp. tauschii]	94.437	
TraesCS5B01G198700.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020169876.1	zinc finger BED domain-containing protein	93.217	

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search hmmer /	CD-Search hmmer /	CD-Search hmmer /	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
								RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]		
TraesCS5B01G199000.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020190052.1	zinc finger BED domain-containing protein RICESLEEPER 3-like [Aegilops tauschii subsp. tauschii]	91.988	
TraesCS5B01G377100.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			ABA94812.1	hAT family dimerisation domain containing protein [Oryza sativa Japonica Group]	58.779	yes
TraesCS5B01G497000.1	1	ZnF_BED	DUF659	Dimer_Tnp_hAT			XP_020173835.1	uncharacterized protein LOC109759423 isoform X2 [Aegilops tauschii subsp. tauschii]	93.814	
TraesCS5B01G501500.1	1	ZnF_BED					XP_020164333.1	protein NLP4-like [Aegilops tauschii subsp. tauschii]	74.965	yes
TraesCS5D01G205800.1	1	ZnF_BED					XP_020169880.1	zinc finger BED domain-containing protein DAYSLEEPER-like [Aegilops tauschii subsp. tauschii]	100	
TraesCS5D01G205900.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020169876.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	98.687	
TraesCS5D01G206100.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020190052.1	zinc finger BED domain-containing protein RICESLEEPER 3-like [Aegilops tauschii subsp. tauschii]	100	
TraesCS5D01G497300.1	1	ZnF_BED	DUF659	Dimer_Tnp_hAT			XP_020173835.1	uncharacterized protein LOC109759423 isoform X2 [Aegilops tauschii subsp. tauschii]	98.836	
TraesCS5D01G501900.1	1	ZnF_BED					XP_020164333.1	protein NLP4-like [Aegilops tauschii subsp. tauschii]	100	yes
TraesCS6A01G049300.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020180515.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	95.421	
TraesCS6A01G115000.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			BAJ88673.1	predicted protein [Hordeum vulgare subsp. vulgare]	87.71	

Gene model	#BED	CD-Search / hmmer	CD-Search / hmmer	CD-Search hmmer /	CD-Search hmmer /	CD-Search hmmer /	Best BLAST hit	Best BLAST hit description	%ID	BED sequence related to BNLs in Neighbour Network Tree
TraesCS6B01G190800.1	1	ZnF_BED	DUF659	Dimer_Tnp_hAT			XP_020182387.1	uncharacterized protein LOC109768065 [Aegilops tauschii subsp. tauschii]	97.734	
TraesCS6B01G423600.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_023157898.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Zea mays]	49.123	
TraesCS6D01G013700.1	3	ZnF_BED	ZnF_BED	ZnF_BED	MISS superfamily NC		PNT68570.1	hypothetical protein BRADI_3g42745v3 [Brachypodium distachyon]	35.981	
TraesCS6D01G103600.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			BAJ88673.1	predicted protein [Hordeum vulgare subsp. vulgare]	89.39	
TraesCS7A01G337100.1	1	ZnF_BED					XP_020168570.1	zinc finger BED domain-containing protein RICESLEEPER 3-like [Aegilops tauschii subsp. tauschii]	98.837	
TraesCS7A01G447400.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_020177565.1	zinc finger BED domain-containing protein RICESLEEPER 2-like [Aegilops tauschii subsp. tauschii]	94.937	yes
TraesCSU01G215500.1	1	ZnF_BED	DUF4413	Dimer_Tnp_hAT			XP_015619043.1	PREDICTED: zinc finger BED domain-containing protein RICESLEEPER 1-like [Oryza sativa Japonica Group]	39.187	

Appendix 8-17. Plant proteomes investigated in section 4.3.6.
Green colour shows proteomes that we selected based the BUSCO analysis and red the ones that did not. We carried out two BUSCO analyses, one with the Viridiplantae ortholog set to allow comparison with algae, and one with the Embryophytes set to refined the analysis on higher plants. Proteomes from plants that are not part of the Embryophytes and scored < 90 % completeness in that analysis but > 90 % completeness in the analysis with the Viridiplantae set were kept. Proteomes from plants that are part of the Embryophytes and scored > 90 % completeness with the Viridiplantae set but < 90 % completeness with the Embryophytes set were no included in the analysis.

Specie	source	%complete	number of busco assessed (viridiplantae_odb10)	%complete	number of busco assessed (embryophyta_odb9)	Kept?	BED domain	BED-NLRs
Aquilegia coerulea	Phytozome	0.99	430.00	0.96	1440		duplication	
Aquilegia coerulea	Phytozome	0.99	430.00	0.96	1440		Y	N
Ananas comosus	Phytozome	0.94	430.00	0.89	1440		Y	N
Actinidia chinensis	Ensembl	0.97	430.00	0.94	1440		Y	N
Aegilops tauschii	http://aegilops.wheat.ucdavis.edu/	0.92	430.00	0.90	1440		Y	Y
Arabidopsis halleri	Phytozome	0.92	430.00	0.86	1440		Y	N
Amaranthus hypochondriacus	Phytozome	0.97	430.00	0.85	1440		Y	N
Arabidopsis lyrata	Phytozome	0.99	430.00	0.99	1440		Y	N
Arabidopsis thaliana	Phytozome	1.00	430.00	0.99	1440		Y	N
Amborella trichopoda	Phytozome	0.97	430.00	0.85	1440			
Brachypodium distachyon	Phytozome	1.00	430.00	0.99	1440		Y	Y
Beta vulgaris	Ensembl	0.99	430.00	0.93	1440		Y	N
Brassica rapa	Phytozome	0.98	430.00	0.97	1440		Y	N
Brassica napus	Ensembl	0.99	430.00	0.97	1440		Y	N
Brassica oleracea	Ensembl	0.97	430.00	0.95	1440		Y	N
Brachypodium stacei	Phytozome	1.00	430.00	0.99	1440		Y	Y
Boechera stricta	Phytozome	0.98	430.00	0.97	1440		Y	N
Citrus clementina	Phytozome	0.98	430.00	0.95	1440		Y	N
Capsella grandiflora	Phytozome	0.97	430.00	0.95	1440		Y	N
Chondrus crispus	Ensembl	0.48	430.00	0.13	1440			
Corchorus capsularis	Ensembl	0.95	430.00	0.86	1440		Y	N
Carica papaya	Phytozome	0.79	430.00	0.72	1440			
Chenopodium quinoa	Phytozome	0.97	430.00	0.91	1440		Y	N
Chlamydomonas reinhardtii	Phytozome	0.97	430.00	0.28	1440		N	N
Capsella rubella	Phytozome	0.98	430.00	0.97	1440		Y	N
Crocus sativus	Phytozome	0.97	430.00	0.89	1440		Y	N

Specie	source	%complete	number of busco assessed (viridiplantae_odb10)	%complete	number of busco assessed (embryophyta_odb9)	Kept?	BED domain	BED-NLRs
Citrus sinensis	Phytozome	0.94	430.00	0.87	1440		Y	N
Coccomyxa subellipsoidea	Phytozome	0.86	430.00	0.27	1440		Y	N
Cyanidioschyzon merolae	Ensembl	0.43	430.00	0.10	1440			
Daucus carota	Phytozome	0.90	430.00	0.83	1440			
Dioscorea rotundata	Ensembl	0.78	430.00	0.68	1440			
Dunaliella salina	Phytozome	0.73	430.00	0.19	1440			
Eucalyptus grandis	Phytozome	0.97	430.00	0.92	1440		Y	Y
Eutrema salsugineum	Phytozome	1.00	430.00	0.98	1440		Y	N
Fragaria vesca	Phytozome	0.94	430.00	0.88	1440		Y	N
Galdieria sulphuraria	Ensembl	0.53	430.00	0.14	1440			
Glycine max	Phytozome	1.00	430.00	0.97	1440		Y	Y
Gossypium raimondii	Phytozome	1.00	430.00	0.97	1440		Y	N
Helianthus annuus	Ensembl	0.98	430.00	0.94	1440		Y	N
Hordeum vulgare	Ensembl	0.92	430.00	0.89	1440		Y	Y
Kalanchoe fedtschenko	Phytozome	0.97	430.00	0.90	1440		Y	N
Kalanchoe laxiflora	Phytozome	1.00	430.00	0.94	1440		Y	N
Leersia perrieri	Ensembl	0.96	430.00	0.94	1440		Y	Y
Lupinus angustifolius	Ensembl	0.97	430.00	0.91	1440		Y	N
Linum usitatissimum	Phytozome	0.99	430.00	0.92	1440		Y	N
Musa acuminata	Phytozome	0.96	430.00	0.87	1440		Y	N
Malus domestica	Phytozome	0.87	430.00	0.87	1440		Y	N
Manihot esculenta	Phytozome	0.99	430.00	0.95	1440		Y	N
Mimulus guttatus	Phytozome	0.99	430.00	0.94	1440		Y	N
Marchantia polymorpha	Phytozome	0.98	430.00	0.67	1440		Y	N
Micromonas pusilla	Phytozome	0.87	430.00	0.21	1440		N	N
Micromonas sp	Phytozome	0.90	430.00	0.23	1440		Y	N
Medicago truncatula	Phytozome	0.98	430.00	0.94	1440		Y	Y
Nicotiana attenuata	Ensembl	0.97	430.00	0.93	1440		Y	N
Ostreococcus lucimarinus	Phytozome	0.83	430.00	0.20	1440			
Oryza brachyantha	Ensembl	0.95	430.00	0.93	1440		Y	N
Oryza sativa indica	Ensembl	0.98	430.00	0.96	1440		Y	Y
Oryza meridionalis	Ensembl	0.86	430.00	0.79	1440			
Oryza sativa japonica	Phytozome	0.97	430.00	0.96	1440		Y	Y

Specie	source	%complete	number of busco assessed (viridiplantae_odb10)	%complete	number of busco assessed (embryophyta_odb9)	Kept?	BED domain	BED-NLRs
Oropetium thomaeum	Phytozome	0.83	430.00	0.70	1440			
Panicum hallii	Phytozome	0.98	430.00	0.97	1440		Y	Y
Physcomitrella patens	Ensembl	0.99	430.00	0.69	1440		Y	N
Physcomitrella patens	Ensembl	0.99	430.00	0.69	1440		Same as above	
Prunus persica	Phytozome	1.00	430.00	0.99	1440		Y	N
Populus trichocarpa	Phytozome	1.00	430.00	0.98	1440		Y	Y
Panicum virgatum	Phytozome	0.97	430.00	0.97	1440		duplication	
Panicum virgatum	Phytozome	0.99	430.00	0.98	1440		Y	Y
Phaseolus vulgaris	Phytozome	0.98	430.00	0.96	1440		duplication	
Phaseolus vulgaris	Phytozome	0.99	430.00	0.96	1440		Y	Y
Ricinus communis	Ensembl	0.96	430.00	0.90	1440		Y	N
Sorghum bicolor	Phytozome	0.99	430.00	0.98	1440		Y	N
Sphagnum fallax	Phytozome	0.99	430.00	0.71	1440		N	N
Setaria italica	Phytozome	0.99	430.00	0.98	1440		Y	Y
Solanum lycopersicum	Phytozome	0.98	430.00	0.96	1440		Y	N
Selaginella moellendorffii	Phytozome	0.94	430.00	0.62	1440		N	N
Spirodela polyrhiza	Phytozome	0.92	430.00	0.80	1440		Y	N
Salix purpurea	Phytozome	1.00	430.00	0.97	1440		Y	Y
Solanum tuberosum	Phytozome	0.84	430.00	0.84	1440			
Setaria viridis	Phytozome	0.99	430.00	0.98	1440		Y	N
Theobroma cacao	Phytozome	1.00	430.00	0.98	1440		Y	N
Trifolium pratense	Phytozome	0.93	430.00	0.88	1440		Y	Y
Triticum dicoccoides	Ensembl	0.98	430.00	0.98	1440		Y	Y
Volvox carteri	Phytozome	0.94	430.00	0.26	1440		Y	N
Vigna angularis	Ensembl	0.90	430.00	0.85	1440			
Vigna adiata	Ensembl	0.89	430.00	0.81	1440			
Vitis vinifera	Ensembl	0.98	430.00	0.96	1440		Y	N
Vitis vinifera	Phytozome	0.97	430.00	0.90	1440		duplication	
Zea_mays	Ensembl	0.97	430.00	0.96	1440		Y	Y
Zostera marina	Phytozome	0.99	430.00	0.85	1440		Y	N
Zea_mays	Phytozome	0.90	430.00	0.92	1440		duplication	

Appendix 8-18. Summary of BED-containing proteins clustering with BED-NLRs in neighbour-net analyses carried out on the BED domain (Figure 4-17, Figure 4-18, Figure 4-19, Figure 4-20, Figure 4-21) on phylogenetic groups defined in Table 4-14. "-" shows non-additional domain so only a single BED domain was identified in the corresponding protein.

Protein ID	Group	Clade	Additional domains		Comment
AET1Gv20001900.2	orange	Yr7 clade	DnaJ	NAM	
AET1Gv20900200.1	orange	Yr7 clade	DUF4413		
AET5Gv20822700.2	orange	Yr7 clade	-		
Bradi4g03540.1.p	orange	Yr7 clade	-		
Bradi4g03620.1.p	orange	Yr7 clade	NAM		
Brast07G116500.1.p	orange	Yr7 clade	-		
Brast10G142600.1.p	orange	Yr7 clade	-		
HORVU3Hr1G025360.1	orange	Yr7 clade	-		
HORVU4Hr1G043750.1	orange	Yr7 clade	-		
HORVU5Hr1G110060.1	orange	Yr7 clade	-		
HORVU6Hr1G010990.2	orange	Yr7 clade	-		
TraesCS1D02G004800.1	orange	Yr7 clade	DnaJ	NAM	
TRIDC6AG015800.1	orange	Yr7 clade	-		
TRIDC6BG073250.1	orange	Clade_I	-		
AET6Gv20991600.1	orange	Clade_I	-		
AET3Gv21007600.1	orange	Clade_I	-		
TraesCS2A02G466500.1	orange	Clade_I	-		
TraesCS2B02G488600.1	orange	Clade_I	-		
AET2Gv21027500.16	orange	Clade_I	-		
TraesCS2B02G488600.1	orange	Clade_II	BED		
AET6Gv20767800.3	orange	Clade_II	-		
Bradi2g25117.1.p	orange	Clade_II	-		
TraesCS4A02G069300.1	orange	Clade_II	-		

Protein ID	Group	Clade	Additional domains			Comment
TraesCS3A02G304000.1	orange	Clade_II	-			
HORVU7Hr1G027110.2	orange	Clade_II	-			
HORVU6Hr1G085090.1	orange	Clade_II	-			
HORVU6Hr1G027260.1	orange	Clade_II	-			
HORVU2Hr1G054220.1	orange	Clade_II	-			
AET2Gv20709500.1	orange	Clade_III	FAM177			
AET2Gv20895600.6	orange	Clade_III	-			
AET4Gv20282600.2	orange	Clade_III	-			
AET5Gv20633400.3	orange	Clade_III	-			
AET5Gv21122100.5	orange	Clade_III	-			
AET6Gv20219900.1	orange	Clade_III	-			
Bradi3g24662.1.p	orange	Clade_III	-			
Brast08G141500.1.p	orange	Clade_III	-			
HORVU5Hr1G114920.5	orange	Clade_III	-			
HORVU7Hr1G057150.1	orange	Clade_III	-			
TraesCS5B02G501500.1	orange	Clade_III	-			Dimer_Tnp_hAT in RefSeqv1.0
TraesCS5D02G501900.1	orange	Clade_III	-			Dimer _Tnp_hAT in RefSeqv1.0
TraesCS7A02G337100.1	orange	Clade_III	-			
TRIDC2AG021560.2	orange	Clade_III	-			
TRIDC5AG069450.1	orange	Clade_III	-			
AET4Gv20131800.3	orange	Clade_IV	DUF4413	Dimer_Tnp_hAT		
AET3Gv20623400.3	orange	Clade_IV	DUF4413	Dimer_Tnp_hAT		
AET5Gv20489600.2	orange	Clade_IV	DUF4413	Dimer_Tnp_hAT	Sec34	
AET2Gv21053100.2	orange	Clade_IV	DUF4413	Dimer_Tnp_hAT	BED	
AET2Gv20258400.6	orange	Clade_IV	DUF4413	Dimer_Tnp_hAT		

Protein ID	Group	Clade	Additional domains		Comment
AET2Gv21053100.2	orange	Clade_IV	DUF4413	Dimer_Tnp_hAT BED	
HORVU0Hr1G031720.1	orange	Clade_IV	-		
HORVU1Hr1G068040.1	orange	Clade_IV	CENP-B_dimeris	Nop14	
HORVU6Hr1G060060.1	orange	Clade_IV	CENP-B_dimeris	Nop14	
HORVU2Hr1G057120.1	orange	Clade_IV	-		
TraesCS5B01G377100.1_1	orange	Clade_IV	DUF4413	Dimer_Tnp_hAT	from RefSeqv1.0 annotation
AET6Gv20294800.1	orange	Yr7 clade	DUF4413	Dimer_Tnp_hAT	
AET4Gv20131600.4	orange	Yr7 clade	DUF4413	Dimer_Tnp_hAT	
Bradi4g03550.1.p	orange	Yr7 clade	-		
Brast10G036100.1.p	orange	Yr7 clade	-		
Bradi1g01593.1	orange	Yr7 clade	-		
BGIOSGA021203-PA	brown	Xa1 clade	-		
BGIOSGA024196-PA	brown	Xa1 clade	DUF4413	Dimer_Tnp_hAT	
BGIOSGA032312-PA	brown	Xa1 clade	-		
BGIOSGA037669-PA	brown	Xa1 clade	-		
BGIOSGA037670-PA	brown	Xa1 clade	-		
BGIOSGA040723-PA	brown	Xa1 clade	-		
LOC_Os02g56290.1	brown	Xa1 clade	DUF659	Dimer_Tnp_hAT	
LOC_Os10g04850.1	brown	Xa1 clade	F-box		
LOC_Os12g23400.1	brown	Xa1 clade	DUF659		
LOC_Os12g39020.1	brown	Xa1 clade	-		
LPERR10G03230.1	brown	Xa1 clade	-		
LPERR12G14140.1	brown	Xa1 clade	-		
BGIOSGA028191-PA	brown	Clade_I	-		
LOC_Os04g22100.1	brown	Clade_I	-		

Protein ID	Group	Clade	Additional domains			Comment
BGIOSGA006434-PA	brown	Clade_II	DUF4413	Dimer_Tnp_hAT		
BGIOSGA006771-PA	brown	Clade_II	DUF4413	Dimer_Tnp_hAT		
BGIOSGA018997-PA	brown	Clade_II	-			
BGIOSGA019405-PA	brown	Clade_II	-			
BGIOSGA037354-PA	brown	Clade_II	DUF4413			
LOC_Os02g15390.1	brown	Clade_II	-			
LOC_Os03g32190.1	brown	Clade_II	-			
LOC_Os11g36790.1	brown	Clade_II	DUF295	Dimer_Tnp_hAT	F-box	
LOC_Os12g16470.1	brown	Clade_II	-			
LPERR12G14130.1	brown	Clade_II	-			
Potri.T026500.1	green	Clade_I	-			
Potri.001G405900.1	green	Clade_I	-			
Potri.T014400.1	green	Clade_I	-			
Potri.T025700.1	green	Clade_I	-			
Potri.T052400.1	green	Clade_I	LRR			misannotated BED-NLR?
Potri.T028400.1	green	Clade_I	-			
Potri.006G031100.1	green	Clade_II	-			
Potri.002G190100.1	green	Clade_II	DUF659			
Potri.006G021300.1	green	Clade_II	DUF659	Dimer_Tnp_hAT	BED	
Potri.009G149800.1	green	Clade_II	DUF659	Dimer_Tnp_hAT		
Potri.013G087500.1	green	Clade_II	DUF659			
Potri.016G018800.1	green	Clade_II	DUF659	Dimer_Tnp_hAT	BED	
Potri.T012100.1	green	Clade_II	-			
SapurV1A.0154s0190.1.p	green	Clade_II	DUF659	Dimer_Tnp_hAT	BED	
SapurV1A.0295s0460.1.p	green	Clade_II	DUF659	Dimer_Tnp_hAT	BED	

Protein ID	Group	Clade	Additional domains			Comment
Potri.001G196200.1	green	Clade_III	DUF659	Dimer_Tnp_hAT		
Potri.T016100.1	green	Clade_III	-			
Potri.014G059400.1	green	Clade_III	DUF659			
Potri.001G192900.1	green	Clade_III	DUF659			
Potri.018G032800.1	green	Clade_III	DUF659	Dimer_Tnp_hAT		
SapurV1A.0070s0660.1.p	green	Clade_IV	DUF659	Dimer_Tnp_hAT		
SapurV1A.0851s0040.1.p	green	Clade_IV	DUF4413	Dimer_Tnp_hAT		
SapurV1A.3382s0020.1.p	green	Clade_IV	DUF4413	Dimer_Tnp_hAT		
Potri.008G011500.1	green	Clade_IV	DUF659	Dimer_Tnp_hAT		
Potri.017G065700.1	green	Clade_IV	-			
Potri.015G030000.1	green	Clade_IV	-			
Potri.018G067100.1	green	Clade_IV	DUF659			
Potri.010G037700.1	green	Clade_IV	DUF659			
Potri.012G049100.1	green	Clade_IV	DUF659			
Potri.012G059200.1	green	Clade_IV	DUF4413	Dimer_Tnp_hAT		
Glyma.06G230700.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.01G121600.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.13G258600.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.02G159600.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.19G103700.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.07G171900.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.17G215100.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.19G239600.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT		
Glyma.10G020400.1.p	blue	Clade_I	DUF659	DUF659	Dimer_Tnp_hAT	
Glyma.07G131600.1.p	blue	Clade_I	SWIB			
Glyma.14G145100.1.p	blue	Clade_I	DUF659	DUF659		

Protein ID	Group	Clade	Additional domains		Comment
Glyma.01G091700.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.11G183000.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.09G105100.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.17G188700.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.15G205800.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.09G064900.1.p	blue	Clade_I	DUF659		
Glyma.08G258600.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.07G163900.1.p	blue	Clade_I	DUF659		
Glyma.16G144300.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.09G104100.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.10G121900.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.19G102600.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.U023300.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.04G108300.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.04G163700.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.18G112000.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.12G152900.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.05G100300.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.18G099100.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.14G162700.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.17G232100.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.13G192700.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.20G072800.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.18G113800.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.20G074600.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.18G118000.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	

Protein ID	Group	Clade	Additional domains		Comment
Glyma.08G294500.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.02G193100.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.04G135600.1.p	blue	Clade_I	DUF659		
Glyma.10G088500.1.p	blue	Clade_I	DUF659		
Glyma.20G079200.1.p	blue	Clade_I	DUF659	Dimer_Tnp_hAT	
Glyma.11G256900.1.p	blue	Clade_I	DUF659		
Tp57577_TGAC_v2_mRNA33150	blue	Clade_I	Dimer_Tnp_hAT		
Tp57577_TGAC_v2_mRNA42047	blue	Clade_I	-		
Tp57577_TGAC_v2_mRNA27081	blue	Clade_II	-		
Glyma.15G148600.1.p	blue	Clade_II	-		
Tp57577_TGAC_v2_mRNA8408	blue	Clade_II	-		
Tp57577_TGAC_v2_mRNA21612	blue	Clade_II	DBD_Tnp_Hermes		
Tp57577_TGAC_v2_mRNA37185	blue	Clade_II	DUF4413	Dimer_Tnp_hAT	
Tp57577_TGAC_v2_mRNA14798	blue	Clade_II	DUF4413		
Tp57577_TGAC_v2_mRNA38505	blue	Clade_II	DUF4413	Dimer_Tnp_hAT	
Tp57577_TGAC_v2_mRNA2311	blue	Clade_II	DUF4413	Dimer_Tnp_hAT	
Eucgr.F00906.1.p	blue	Clade_III	BED		
Eucgr.L01859.1.p	blue	Clade_III	BED		
Eucgr.F00547.1.p	blue	Clade_III	BED		
Eucgr.F00840.1.p	blue	Clade_III	BED		
Tp57577_TGAC_v2_mRNA27203	blue	Clade_III	DUF659	Dimer_Tnp_hAT	
Glyma.08G154000.1.p	blue	Clade_III	DUF659	Dimer_Tnp_hAT	
Glyma.04G117600.1.p	blue	Clade_III	DUF659	Dimer_Tnp_hAT	
Glyma.01G096900.1.p	blue	Clade_III	DUF659	DUF659	
Glyma.19G056600.1.p	blue	Clade_IV	DUF659		
Glyma.17G188400.1.p	blue	Clade_IV	DUF659		

Protein ID	Group	Clade	Additional domains		Comment
Tp57577_TGAC_v2_mRNA27077	blue	Clade_IV	-		
Zm00001d051141_P001	yellow	Clade_I	-		
Zm00001d032348_P001	yellow	Clade_I	-		
Zm00001d015593_P001	yellow	Clade_I	-		
Zm00001d025207_P001	yellow	Clade_I	-		
Zm00001d043354_P006	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Zm00001d006692_P001	yellow	Clade_I	Myb_DNA-bind		
Zm00001d040607_P001	yellow	Clade_I	-		
Zm00001d034670_P002	yellow	Clade_I	DUF659	Dimer_Tnp_hAT	
Zm00001d023717_P001	yellow	Clade_I	-		
Seita.1G225700.1.p	yellow	Clade_I	DUF1342		
Seita.6G147100.1.p	yellow	Clade_I	-		
Pahal.B02480.1	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Pahal.I01262.1	yellow	Clade_I	DUF659	Dimer_Tnp_hAT	
Pahal.D02150.1	yellow	Clade_I	DUF4413		
Pahal.B01450.1	yellow	Clade_I	-		
Pahal.B03206.1	yellow	Clade_I	-		
Pahal.A02558.1	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Pahal.G00842.1	yellow	Clade_I	-		
Pahal.I04669.1	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Pahal.F02765.1	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Pahal.F02209.1	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Pavir.4KG100900.1.p	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Pavir.5KG119600.1.p	yellow	Clade_I	-		
Pavir.7NG325900.1.p	yellow	Clade_I	-		
Pavir.7NG440300.1.p	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	

Protein ID	Group	Clade	Additional domains		Comment
Pavir.5NG489100.1.p	yellow	Clade_I	-		
Pavir.5KG219700.1.p	yellow	Clade_I	-		
Pavir.6KG164700.1.p	yellow	Clade_I	-		
Pavir.9KG442200.1.p	yellow	Clade_I	-		
Pavir.6NG184800.1.p	yellow	Clade_I	-		
Pavir.5KG333700.1.p	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	
Pavir.4KG097700.1.p	yellow	Clade_I	-		
Pavir.7NG282300.1.p	yellow	Clade_I	-		
Pavir.5KG731000.1.p	yellow	Clade_I	-		
Pavir.2KG096500.1.p	yellow	Clade_I	-		
Pavir.2KG516200.1.p	yellow	Clade_I	-		
Pavir.2NG560200.1.p	yellow	Clade_I	-		
Pavir.3KG240400.1.p	yellow	Clade_I	-		
Pavir.4NG167500.1.p	yellow	Clade_I	-		
Pavir.6KG348400.1.p	yellow	Clade_I	-		
Pavir.9KG245400.1.p	yellow	Clade_I	-		
Pavir.2KG362800.1.p	yellow	Clade_I	-		
Pavir.5NG487100.1.p	yellow	Clade_I	-		
Pavir.9NG336800.1.p	yellow	Clade_I	-		
Pavir.1NG092200.1.p	yellow	Clade_I	-		
Pavir.2NG178900.1.p	yellow	Clade_I	-		
Pavir.3KG155900.1.p	yellow	Clade_I	-		
Pavir.J445200.1.p	yellow	Clade_I	-		
Pavir.5NG287600.1.p	yellow	Clade_I	-		
Pavir.2KG194600.1.p	yellow	Clade_I	-		
Pavir.7NG041700.1.p	yellow	Clade_I	-		

Protein ID	Group	Clade	Additional domains					Comment
Pavir.9NG118300.1.p	yellow	Clade_I	-					
Pavir.4NG021300.1.p	yellow	Clade_I	-					
Pavir.J080500.1.p	yellow	Clade_I	-					
Pavir.5NG156000.1.p	yellow	Clade_I	-					
Pavir.9KG062400.1.p	yellow	Clade_I	DUF4413	Dimer_Tnp_hAT	NPR1_like_C	Ank	DUF3420	
Pavir.2KG112400.1.p	yellow	Clade_I	-					
Pavir.J361500.1.p	yellow	Clade_I	-					
Pavir.1NG454900.1.p	yellow	Clade_I	-					
Pavir.8KG178700.1.p	yellow	Clade_I	-					
Pavir.3KG548500.1.p	yellow	Clade_I	-					
Pavir.9KG258000.1.p	yellow	Clade_I	-					
Pavir.6KG210100.1.p	yellow	Clade_II	DUF4413	Dimer_Tnp_hAT				
Pavir.3KG398100.1.p	yellow	Clade_II	BED					
Pavir.6NG334600.1.p	yellow	Clade_II	BED	NAM				
Pavir.9KG517200.1.p	yellow	Clade_II	-					
Pavir.9NG821600.1.p	yellow	Clade_II	-					
Pavir.8NG205900.1.p	yellow	Clade_II	DUF4413					
Pavir.J650000.1.p	yellow	Clade_II	-					
Pavir.6KG023200.1.p	yellow	Clade_II	-					
Pahal.F00101.1	yellow	Clade_II	-					
Pahal.F01374.1	yellow	Clade_II	-					
Pahal.C03786.1	yellow	Clade_II	BED					
Seita.7G242000.1.p	yellow	Clade_II	-					
Zm00001d003194_P001	yellow	Clade_II	DUF4413					
Pavir.1KG445100.1.p	yellow	Clade_III	-					
Pavir.5KG495100.1.p	yellow	Clade_III	DUF4413	Dimer_Tnp_hAT				

Protein ID	Group	Clade	Additional domains				Comment
Pavir.2NG217300.1.p	yellow	Clade_III	DUF4413	Dimer_Tnp_hAT			
Pavir.9NG116000.1.p	yellow	Clade_III	DUF4413	Dimer_Tnp_hAT			
Pavir.1NG316100.1.p	yellow	Clade_III	Dimer_Tnp_hAT				
Pavir.8KG229800.1.p	yellow	Clade_III	-				
Pavir.3KG343900.1.p	yellow	Clade_III	-				
Pavir.9NG342800.1.p	yellow	Clade_III	Dimer_Tnp_hAT				
Pavir.3KG233300.1.p	yellow	Clade_III	-				
Pavir.6NG050600.1.p	yellow	Clade_III	DUF4413	Dimer_Tnp_hAT			
Pavir.9NG394500.1.p	yellow	Clade_III	Dimer_Tnp_hAT				
Pavir.6KG041000.1.p	yellow	Clade_III	-				
Pavir.5KG154400.1.p	yellow	Clade_III	DUF4413	Dimer_Tnp_hAT			
Pavir.9KG159700.1.p	yellow	Clade_III	Fbox				
Pavir.5NG476500.1.p	yellow	Clade_III	Fbox				
Pahal.C03699.1	yellow	Clade_III	-				
Pahal.J00339.1	yellow	Clade_III	-				
Pahal.E02830.1	yellow	Clade_III	Fbox				
Seita.5G177200.1.p	yellow	Clade_III	Fbox				
Zm00001d013336_P002	yellow	Clade_III	DUF4413	DUF4413	Dimer_Tnp_hAT	BED	
Zm00001d033903_P001	yellow	Clade_III	Tmemb_14				
Zm00001d013336_P002	yellow	Clade_III	DUF4413	DUF4413	Dimer_Tnp_hAT	BED	
Zm00001d022534_P001	yellow	Clade_III	-				

Appendix 8-19. List of primers used for Sanger Sequencing to verify sequences of the *Yr7* cassette for wheat transformation.
L stands for the left border and R for the right border of the PCR product.

Forward primer name	Forward primer sequence	Reverse primer name	Reverse primer sequence	Product size (bp)	Comment
F_flanking_insert_pICH47742	A_Sr33P_R	GAACCCTGTGGTTGGCATGCACATAC	AGAAAGATGGGAGGGAAA	901	Level 1 construct
B_Sr33P_L	B_Sr33P_R	TGCTGATCTCATCCATCC	GGACAACGAAGCGAAAG	1024	Level 1 construct
C_Sr33P_L	C_Yr7_R	TGTCCCCTCACGGATCT	AGACTGGAGCCTCTCGAC	997	Level 1 construct + T1 and T2 genotyping
D_Yr7_L	D_Yr7_R	CTGGTGCAGACCATCCT	GGCTCGAGAACTTTACTCA	991	Level 1 construct
E_Yr7_L	E_Yr7_R	AGACCACCGCAGCTAAC	ACGGTACGTCAACTCATCA	980	Level 1 construct
F_Yr7_L	F_Yr7_R	GAAATCGCTGGGACTCA	CTCGTTCATGGATTGGAG	985	Level 1 construct
G_Yr7_L	G_Yr7_R	TCCAGGAGCTACATGATTT	AAGATGACTGAAACCTTCG	980	Level 1 construct
H_Yr7_L	H_Yr7_R	AAAACTTTGGCGTTCCAT	GGGTCATTGATGTCAAGC	992	Level 1 construct
I_Yr7_L	I_Sr33T_R	TCAGACTGTCCTGGCTTG	CTCAGACGCCACTAGCAG	993	Level 1 construct
J_Sr33T_L	J_Sr33T_R	TGGTAAAGTTGCATTTTGG	TAGAATTTGGGCTTCATTT	870	Level 1 construct
K_Sr33T_L	R_ flanking_insert_pICH47742	TTCTGCCAATGTGTTTCC	CTGGTGGCAGGATATATTGTGGTG	348	Level 1 construct
Plasmid/Resistance-cassette_L	Plasmid/Resistance-cassette_R	AAGCCTGCGAAGAGTTG	ATCAGAGCTTGGTTGACG	816	Level 2 construct
Resistance-Cassette_Sr33P_L	Resistance-Cassette_Sr33P_R	TGTGAGAATTCGCCTGAA	TGGAAACAAATCGACAGG	814	Level 2 construct
Sr33T-linker-plasmid_L	Sr33T-linker-plasmid_R	TTCTGCCAATGTGTTTCC	CATCTGTCAGCACTCTGC	811	Level 2 construct

Appendix 8-20. List of primer used to generate the *Yr7* CDS construct and *Yr7* CDS carrying the D646V mutation in MHD motif. Primers contain tails with the *BpiI* restriction sites and 4 bp overhangs for cloning in pUAP1 level 0 acceptor via Golden Gate cloning. Primers also contain tails with deletion of the STOP codon for protein C-terminus tagging in the level 1 constructs.

Name	Forward primer	Reverse primer	Forward primer sequence	Reverse primer sequence	product size	Comment
exon1	F_rm_ATG_pUAP1	7_E1_R	ATGAAGACGTCTCAAATGGAGCCGGCGGGAGACT	TTGAAGACATCTCCTTCGACTTGGTGTTG	337 bp	Yr7 CDS
exon2	7_E1-2_F	7_E2_R	ATGAAGACAAGGAGTTACAAGTGACGAGCCTGAC	TTGAAGACATCTTGAAGTGTTTGGGGGGTG	297 bp	Yr7 CDS
exon3	7_E2-3_F	STOP_pUAP1	ATGAAGACAACAAGCACCGGCGATGCTACTTG	ATGAAGACGTCTCGAAGCTTAATTCACATATCGACCA	4.2 kb	Yr7 CDS
exon3-no-STOP	7_E2-3_F	oh_tag_pUAP1	ATGAAGACAACAAGCACCGGCGATGCTACTTG	ATGAAGACGTCTCGCGAATTCACATATCGACCATCAATTTTGA	4.2 kb	Yr7 CDS + tag
MHV_fragment 1	F_rm_ATG_pUAP1	7_MHV_R_1	ATGAAGACGTCTCAAATGGAGCCGGCGGGAGACT	TTGAAGACATAACATGCATGAGATCACACAT	2.1 kb	D646V mutation (in MHD motif)
MHV_fragment 2	7_MHV_F_2	oh_tag_pUAP1	ATGAAGACAATGTTTTCGCAAGGATGATTTC	ATGAAGACGTCTCGCGAATTCACATATCGACCATCAATTTTGA	2.8 kb	D646V mutation (in MHD motif)

Appendix 8-21. List of primer used to generate the truncations in the *Yr7* CDS construct. Primers contain tails with the *BpiI* restriction sites and 4 bp overhangs for cloning in pUAP1 level 0 acceptor via Golden Gate cloning. Primers also contain tails with deletion of the STOP codon for protein C-terminus tagging in the level 1 constructs. Constructs highlighted in orange are derived from synthesized *Yr7* codon-optimised for expression in *N. benthamiana*.

Name	Forward primer	Reverse primer	Forward primer sequence	Reverse primer sequence	product size	Comment
Yr7-exon1	F_rm_ATG_pUAP1	Yr7-E1-ohtag-R	ATGAAGACGTCTCAAATGGAGCCGGCGGGAGACT	TGAAGACGTCTCGCGAAGCTCCTTCGACTTGGTGTTGGAG	337	Yr7-Exon1-Truncation
Yr7-exon2	Yr7-E2-ohtag-F	Yr7-E2-ohtag-R	ATGAAGACGTCTCAAATGACAAGTGACGAGCCTGACG	ATGAAGACGTCTCGCGAAGCTGAAGTGTTTGGGGGGT	297	Yr7-Exon2-Truncation
Yr7-BED	Yr7-BED-ohprom_F	Yr7-BED-ohtag_R	ATGAAGACGTCTCAAATGTCCCCGGTATGGGAACA	ATGAAGACGTCTCGCGAAGCGGAATGCTCCTTCTCCAA	191	Yr7-BED-Truncation
Yr7-AA201	F_rm_ATG_pUAP1	Yr7-E2-ohtag-R	ATGAAGACGTCTCAAATGGAGCCGGCGGGAGACT	ATGAAGACGTCTCGCGAAGCTGAAGTGTTTGGGGGGT	638	Yr7-AA201-Truncation
Yr7-AA-242	Yr7-opt-E1-F_F	Yr7-AA-242_R	ATGAAGACGTCTCAAATGGAACCCGCTGGAGATTC	ATGAAGACGTCTCGCGAACACTGAGCGTGAATGTT	765	Yr7-AA242-Truncation
Yr7-AA-308	Yr7-opt-E1-F_F	Yr7-AA-308_R	ATGAAGACGTCTCAAATGGAACCCGCTGGAGATTC	ATGAAGACGTCTCGCGAACTACCGTACACTTCCAT	954	Yr7-AA308-Truncation
Yr7-dNLS-fragment_1	Yr7-opt-E1-F_F	Yr7_R_NLS_F1	ATGAAGACGTCTCAAATGGAACCCGCTGGAGATTC	ATGAAGACGTGATCAGAAGAGCCCACCTCC	654	NLS deletion mutant
Yr7-dNLS-fragment_2_AA242	Yr7_F_NLS_F2	Yr7-AA-242_R	ATGAAGACGTGATCCAACACAGACTACT	ATGAAGACGTCTCGCGAACACTGAGCGTGAATGTT	126	NLS deletion mutant
Yr7-dNLS-fragment_2_AA308	Yr7_F_NLS_F2	Yr7-AA-308_R	ATGAAGACGTGATCCAACACAGACTACT	ATGAAGACGTCTCGCGAACTACCGTACACTTCCAT	297	NLS deletion mutant

Appendix 8-22. List of the different transcriptional units including regulatory elements and protein tags that were used transient expression in *N. benthamiana*

Constructs	Promoter	Synbio/Addgene #	Terminator	Synbio/Addgene #	Tag	Synbio/Addgene #	Experiment
Yr7-CDS-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-CDS-YFP	35S	pICH51266/#50267	35S	pICH51266/#50267	YFP tag (yellow variant of GFP)	pICSL50005	Western blot and cellular localization
D646V-Yr7-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-exon1-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-exon2-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-BED-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-BED-YFP	35S	pICH51266/#50267	35S	pICH51266/#50267	YFP tag (yellow variant of GFP)	pICSL50005	Western blot and cellular localization
Yr7-AA201-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot

Constructs	Promoter	Synbio/Addgene #	Terminator	Synbio/Addgene #	Tag	Synbio/Addgene #	Experiment
Yr7-AA201-YFP	35S	pICH51266/#50267	35S	pICH51266/#50267	YFP tag (yellow variant of GFP)	pICSL50005	Western blot and cellular localization
Yr7-AA-242-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-AA-242-YFP	35S	pICH51266/#50267	35S	pICH51266/#50267	YFP tag (yellow variant of GFP)	pICSL50005	Western blot and cellular localization
Yr7-AA-308-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-AA-308-YFP	35S	pICH51266/#50267	35S	pICH51266/#50267	YFP tag (yellow variant of GFP)	pICSL50005	Western blot and cellular localization
Yr7-dNLS-AA-242-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-dNLS-AA-242-YFP	35S	pICH51266/#50267	35S	pICH51266/#50267	YFP tag (yellow variant of GFP)	pICSL50005	Western blot and cellular localization
Yr7-dNLS-AA-308-HA	35S	pICH51266/#50267	35S	pICH51266/#50267	6xHA (Human influenza hemagglutinin)	pICSL50009A	Western blot
Yr7-dNLS-AA-308-YFP	35S	pICH51266/#50267	35S	pICH51266/#50267	YFP tag (yellow variant of GFP)	pICSL50005	Western blot and cellular localization

Appendix 8-23

Clemence Marchal, Jianping Zhang, Peng Zhang, Paul Fenwick, Burkhard Steuernagel, Nikolai M. Adamski, Lesley Boyd, Robert McIntosh, Brande B. H. Wulff, Simon Berry, Evans Lagudah & Cristobal Uauy. BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust. Nature Plants 4, 662–668 (2018). DOI: 10.1038/s41477-018-0236-4

nature plants

LETTERS

https://doi.org/10.1038/s41477-018-0236-4

BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust

Clemence Marchal[1,7], Jianping Zhang[2,3,4,7], Peng Zhang[2], Paul Fenwick[5], Burkhard Steuernagel[1], Nikolai M. Adamski[1], Lesley Boyd[6], Robert McIntosh[2], Brande B. H. Wulff[1], Simon Berry[5], Evans Lagudah[3] and Cristobal Uauy[1*]

Crop diseases reduce wheat yields by ~25% globally and thus pose a major threat to global food security[1]. Genetic resistance can reduce crop losses in the field and can be selected through the use of molecular markers. However, genetic resistance often breaks down following changes in pathogen virulence, as experienced with the wheat yellow (stripe) rust fungus *Puccinia striiformis* f. sp. *tritici* (*Pst*)[2]. This highlights the need to (1) identify genes that, alone or in combination, provide broad-spectrum resistance, and (2) increase our understanding of the underlying molecular modes of action. Here we report the isolation and characterization of three major yellow rust resistance genes (*Yr7*, *Yr5* and *YrSP*) from hexaploid wheat (*Triticum aestivum*), each having a distinct recognition specificity. We show that *Yr5*, which remains effective to a broad range of *Pst* isolates worldwide, is closely related yet distinct from *Yr7*, whereas *YrSP* is a truncated version of *Yr5* with 99.8% sequence identity. All three *Yr* genes belong to a complex resistance gene cluster on chromosome 2B encoding nucleotide-binding and leucine-rich repeat proteins (NLRs) with a non-canonical N-terminal zinc-finger BED domain[3] that is distinct from those found in non-NLR wheat proteins. We developed diagnostic markers to accelerate haplotype analysis and for marker-assisted selection to expedite the stacking of the non-allelic *Yr* genes. Our results provide evidence that the BED-NLR gene architecture can provide effective field-based resistance to important fungal diseases such as wheat yellow rust.

In plant immunity, NLRs act as intracellular immune receptors that on pathogen recognition trigger a series of signalling steps that ultimately lead to cell death, thus preventing the spread of infection[4,5]. The NB-ARC domain is the hallmark of NLRs, which in most cases include leucine-rich repeats (LRRs) at the C-terminus. Recent in silico analyses have identified NLRs with additional 'integrated' domains[6–8], including zinc-finger BED domains (BED-NLRs). The BED-domain function within BED-NLRs is unknown, although the BED domain from the non-NLR DAYSLEEPER protein was shown to bind DNA in *Arabidopsis*[9]. BED-NLRs are widespread across angiosperm genomes[6–8] and this gene architecture has been shown to confer resistance to bacterial blight in rice (*Xa1*[10,11]).

The genetic relationship between *Yr5* and *Yr7* has been debated for almost 45 years[12,13]. Both genes map to chromosome arm 2BL in hexaploid wheat and were hypothesized to be allelic[14], and closely linked with *YrSP*[15]. While only two of >6,000 tested *Pst* isolates worldwide have been found virulent to *Yr5* (Supplementary Table 1[16,17]), both *Yr7* and *YrSP* have been overcome in the field. For *Yr7*, this is probably due to its wide deployment in cultivars (Supplementary Table 2 and Supplementary Fig. 1). This highlights the importance of stewardship plans (including diagnostic markers) to deploy *Yr5* in combination with other genes as currently done in the USA (for example, *Yr5* + *Yr15*; UC Davis breeding programme).

To clone the genes encoding *Yr7*, *Yr5* and *YrSP*, we identified susceptible ethyl methanesulfonate-derived (EMS) mutants from different genetic backgrounds carrying these genes (Fig. 1 and Supplementary Tables 3 and 4). We performed MutRenSeq[18] and isolated a single candidate contig for each of the three genes based on nine, ten, and four independent susceptible mutants, respectively (Fig. 1a and Supplementary Fig. 2). The three candidate contigs were genetically linked to a common mapping interval, previously identified for the three *Yr* loci[15,19,20]. No recombinant was previously found between *Yr7* and *Yr5* among 143 F_3 progenies[14] and we observed no recombination between *YrSP* and *Yr7* (208 F_3 lines) nor *YrSP* and *Yr5* (256 F_3 lines; Supplementary Table 5). Their closest homologues in the Chinese Spring wheat genome sequence (RefSeq v1.0) all lie within this common genetic interval (Fig. 1b; Supplementary Fig. 3).

Within each contig we predicted a single open reading frame based on RNA-Seq data. All three predicted *Yr* genes displayed similar exon-intron structures (Fig. 1a), although *YrSP* was truncated in exon 3 due to a single base deletion that resulted in a premature termination codon. The 23 mutations identified by MutRenSeq were confirmed by Sanger sequencing and all lead to either an amino-acid substitution or a truncation allele (splice junction or termination codon) (Fig. 1a and Supplementary Table 4). The DNA sequences of *Yr7* and *Yr5* were 77.9% identical across the complete gene, whereas *YrSP* was a truncated version of *Yr5*, sharing 99.8% identity in the common sequence (Supplementary Files 1 and 2). This high sequence identity between *YrSP* and *Yr5* is on a par with that seen for previously characterized allelic series in the wheat *Pm3* (>97% identity)[21] and flax *L* (>90% identity)[22] resistance genes and would suggest that *Yr5* and *YrSP* are allelic. Based on this evidence, we cannot discard the alternative explanations that *Yr5* and *YrSP* are closely linked paralogous genes that arose from a very recent duplication event or that *Yr7* is an allele of *Yr5* that originated from a very diverse haplotype. The absence of recombination between the pairwise populations suggests that *Yr7*, *Yr5* and *YrSp* are linked in repulsion, but we cannot discriminate between paralogous or

[1]John Innes Centre, Norwich Research Park, Norwich, UK. [2]University of Sydney, Plant Breeding Institute, Cobbitty, New South Wales, Australia. [3]Commonwealth Scientific and Industrial Research Organization (CSIRO) Agriculture & Food, Canberra, Australian Capital Territory, Australia. [4]Henan Tianmin Seed Company Limited, Lankao County, Henan Province, China. [5]Limagrain UK Ltd, Rothwell, Market Rasen, Lincolnshire, UK. [6]National Institute of Agricultural Botany (NIAB), Cambridge, UK. [7]These authors contributed equally: C. Marchal and J. Zhang. *e-mail: cristobal.uauy@jic.ac.uk

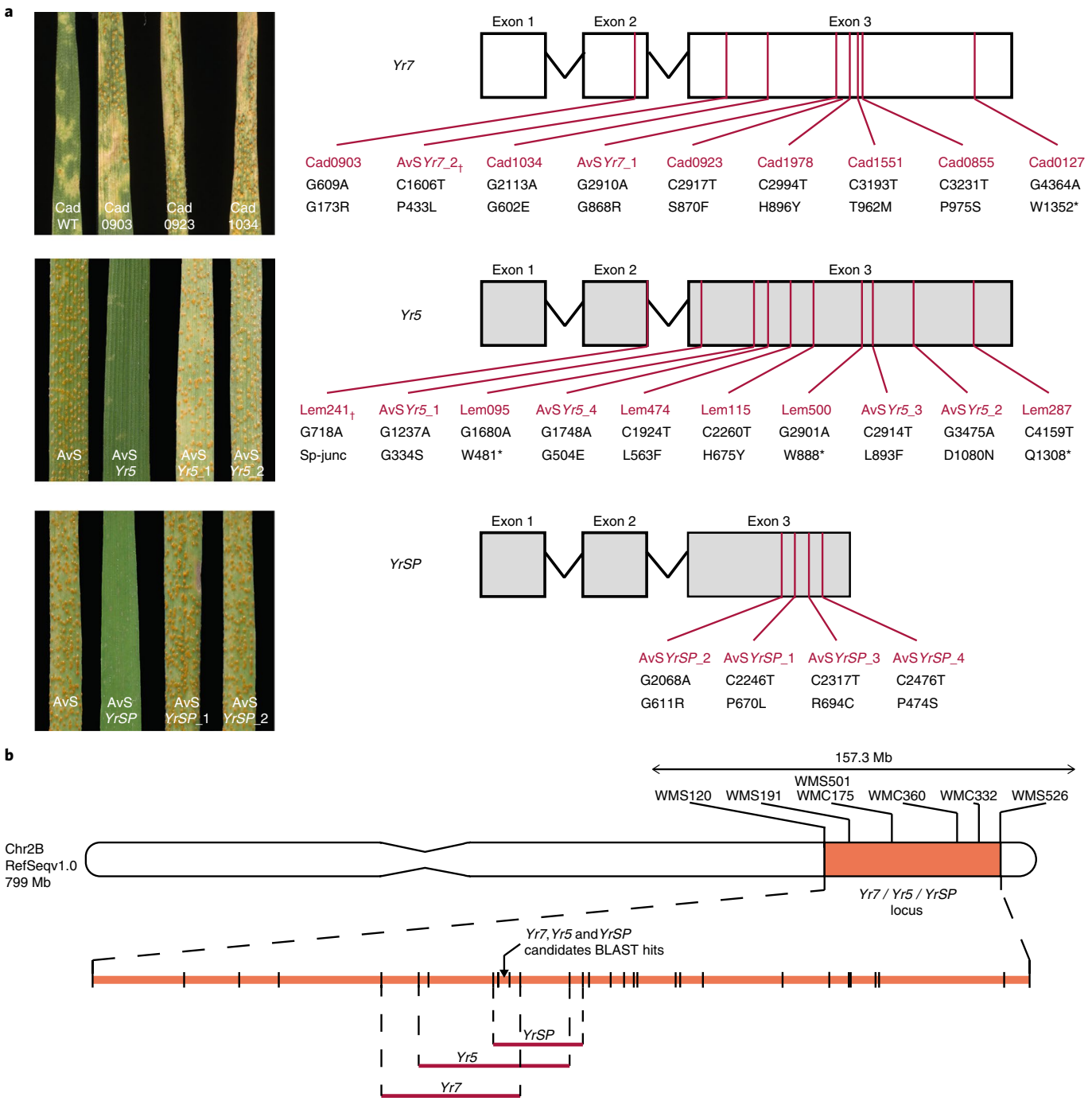


Fig. 1 | *Yr5* and *YrSP* are closely related sequences and distinct from *Yr7*. a, Left: wild-type and selected EMS-derived susceptible mutant lines for *Yr7* (top), *Yr5* (middle) and *YrSP* (bottom) (Supplementary Tables 3 and 4) inoculated with *Pst* isolate 08/21 (*Yr7*), *Pst* 150 E16A + (*Yr5*), or *Pst* 134 E16A + (*YrSP*). Inoculations were performed independently at least three times per mutant line. Right: candidate gene structures, with mutations in red, and their predicted effects on the translated protein. Crosses show mutations shared by two independent mutant lines (Supplementary Table 4). **b**, Schematic representation of the physical interval of the *Yr* loci. The *Yr7*/*Yr5*/*YrSP* locus is shown in orange on chromosome 2B with previously published SSR markers in black. Markers developed in this study to confirm the genetic linkage between the phenotype and the candidate contigs are shown as black vertical lines in the expanded 157.3 Mb interval. *Yr* loci mapping intervals are defined by the red horizontal lines below the expanded chromosome. A more detailed genetic map is shown in Supplementary Fig. 3.

allelic relationships. However, the high sequence identity alongside the genetic analyses support the hypothesis that *Yr5* and *YrSP* are derived from a common sequence and most probably constitute alleles, whereas *Yr7* is encoded by a closely related, yet distinct, gene.

The *Yr7*, *Yr5* and *YrSP* proteins contain a zinc-finger BED domain at the N-terminus, followed by the canonical NB-ARC domain. Unlike previously cloned resistance genes in grasses (such as *Mla10*[23], *Sr33*[24], *Pm3*[25]), neither *Yr7* nor *Yr5*/*YrSP* encode Coiled Coil domains at the N-terminus (Supplementary Fig. 4). Only the *Yr7* and *Yr5* proteins encode multiple LRR motifs at the C-terminus (Fig. 2a; green bars), *YrSP* having lost most of the LRR region due to the premature termination codon in exon 3. *YrSP* still confers

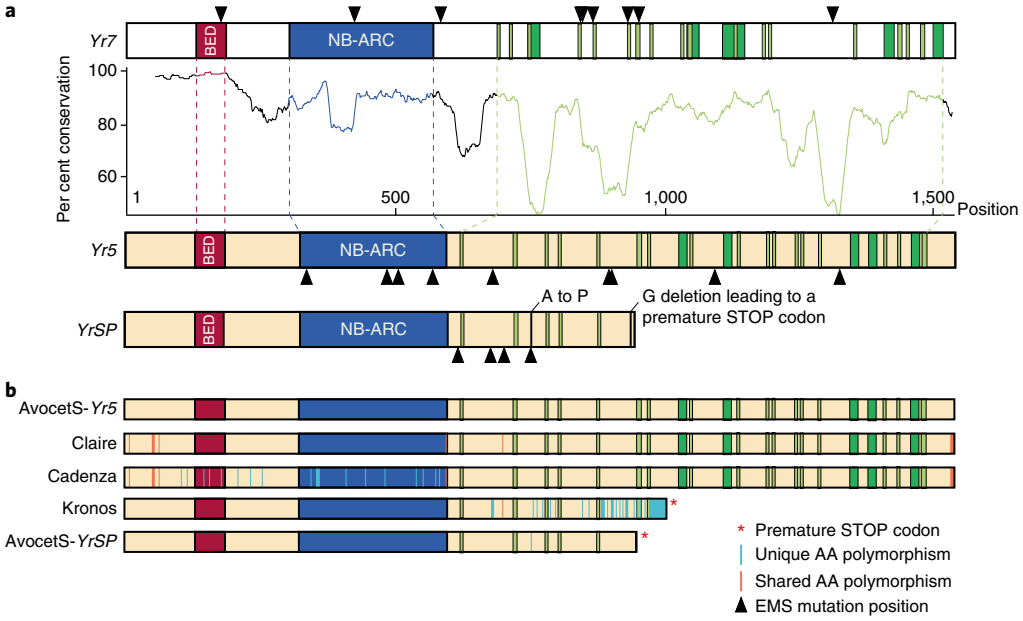


Fig. 2 | *Yr7* and *Yr5/YrSP* encode integrated BED-domain immune receptor genes a, Schematic representation of the *Yr7*, *Yr5* and *YrSP* protein domain organization. BED domains are highlighted in red, NB-ARC domains are in blue, LRR motifs from NLR-Annotator are in dark green, and manually annotated LRR motifs (xxLxLxx) are in light green. Black triangles represent the EMS-induced mutations within the protein sequence. The plot shows the degree of amino-acid conservation (50 amino-acid rolling average) between *Yr7* and *Yr5* proteins, based on the conservation diagram produced by Jalview (2.10.1) from the protein alignment. Regions that correspond to the conserved domains have matching colours. The amino-acid changes between *Yr5* and *YrSP* are annotated in black on the *YrSP* protein. **b**, Five *Yr5/YrSP* haplotypes were identified in this study. Polymorphisms are highlighted across the protein sequence with orange vertical bars for polymorphisms shared by at least two haplotypes and blue vertical bars for polymorphisms that are unique to the corresponding haplotype. Matching colours across protein structures illustrate 100% sequence conservation.

functional resistance to *Pst*, although with a recognition specificity different from *Yr5* (Supplementary Table 1; all isolates virulent to *YrSP* are avirulent to *Yr5*, whereas the two isolates virulent to *Yr5* are avirulent to *YrSP*[16]). *Yr7* and *Yr5/YrSP* are highly conserved in the N-terminus, with a single amino-acid change in the BED domain. This high degree of conservation is eroded downstream of the BED domain (Fig. 2a). The BED domain is required for *Yr7*-mediated resistance, as a single amino-acid change in mutant line Cad0903 led to a susceptible reaction (Fig. 1a). However, recognition specificity is not solely governed by the BED domain, as *Yr5* and *YrSP* have identical BED-domain sequences, yet confer resistance to different *Pst* isolates. The highly conserved *Yr7* and *Yr5/YrSP* BED domains could function in a similar way to the integrated WRKY domain in the *Arabidopsis* RRS1-R immune receptor, which binds unrelated bacterial effectors yet activates a defence response through mechanisms involving other regions of the protein[26].

We examined the variation in *Yr7*, *Yr5* and *YrSP* across eight sequenced tetraploid and hexaploid wheat genomes (Supplementary Table 6). We identified *Yr7* only in Cadenza and Paragon, which are identical-by-descent in this interval (Supplementary File 3, Supplementary Table 7, and Supplementary Fig. 5). Both cultivars are derived from the original source of *Yr7*, tetraploid durum wheat (*T. turgidum* ssp. *durum*) cultivar Iumillo and its hexaploid derivative Thatcher (Supplementary Fig. 5). None of the three sequenced tetraploid accessions (Svevo, Kronos, Zavitan) carries *Yr7* (Supplementary Table 7).

For *Yr5/YrSP*, we identified three additional haplotypes in the sequenced hexaploid wheat cultivars (Fig. 2b and Supplementary Table 8). Cultivar Claire encodes a complete NLR with six amino-acid changes, including one within the NB-ARC domain, and six polymorphisms in the C-terminus compared with *Yr5*. Cultivars Robigus, Paragon, and Cadenza also encode a full-length NLR that shares common polymorphisms with Claire, in addition to 19 amino-acid substitutions across the BED and NB-ARC domains. The presence of the *Yr5/YrSP* haplotype in Cadenza (which also carries *Yr7*) further supports the non-allelic relationship of these genes. The C-terminus polymorphisms between *Yr5* and the other cultivars is due to a 774 bp insertion in *Yr5*, close to the 3' end, which carries an alternate termination codon (Supplementary File 2). Tetraploid cultivars Kronos and Svevo encode a fifth *Yr5/YrSP* haplotype with a truncation in the LRR region distinct from *YrSP*, in addition to multiple amino-acid substitutions across the C-terminus (Supplementary Table 8). This truncated tetraploid haplotype is reminiscent of *YrSP* and is expressed in Kronos (see Methods). However, none of these cultivars (Claire, Robigus, Paragon, Cadenza, Svevo or Kronos) exhibits a *Yr5/YrSP* resistance response, suggesting that these amino-acid changes and truncations may alter recognition specificity or protein function. Additional testing of these haplotypes will provide insight into whether they represent a functional allelic series.

We designed diagnostic markers for *Yr7*, *Yr5* and *YrSP* to facilitate their detection and use in breeding. We confirmed their presence in the donor cultivars Thatcher and Lee (*Yr7*), Spaldings Prolific (*YrSP*), and spelt wheat cv. Album (*Yr5*) (Supplementary Tables 9 and 10 and Supplementary Figs. 5 and 6). We tested the *Yr7* and *YrSP* markers in a collection of global landraces[27] and European cultivars[28] released over the past century. *YrSP* was absent from the tested germplasm, except for AvocetS-*YrSP* (Supplementary Table 10). On the other hand *Yr7* was more prevalent in the germplasm tested and we could track its presence across pedigrees, including Cadenza-derived cultivars (Supplementary Tables 9 and 10 and Supplementary Fig. 5). We confirmed *Yr5* in the AvocetS-*Yr5* and Lemhi-*Yr5* lines, in addition to wheat cultivars in which *Yr5* has been introduced, using gel-based flanking markers (Supplementary Table 11 and Supplementary Fig. 6). The *Yr5* diagnostic marker will facilitate its deployment, hopefully within a breeding strategy that ensures its effectiveness in the long term[29].

a

Syntenic interval	No NLRs	No BED NLRs	No BED NLRs-I	No BED NLRs-II	No BED NLRs-I-II
Rice	6	2	–	–	–
Brachypodium	4	4	1	1	–
Barley	2	–	–	–	–
Aegilops tauschii (D)	8	4	1	1	1
Hexaploid wheat (D)	6	2	1	–	1
Tetraploid wheat (A)	8	1	1	–	–
Hexaploid wheat (A)	12	5	3	2	–
Tetraploid wheat (B)	20	10	6	2	1
Hexaploid wheat (B)	13	5	1	1	3

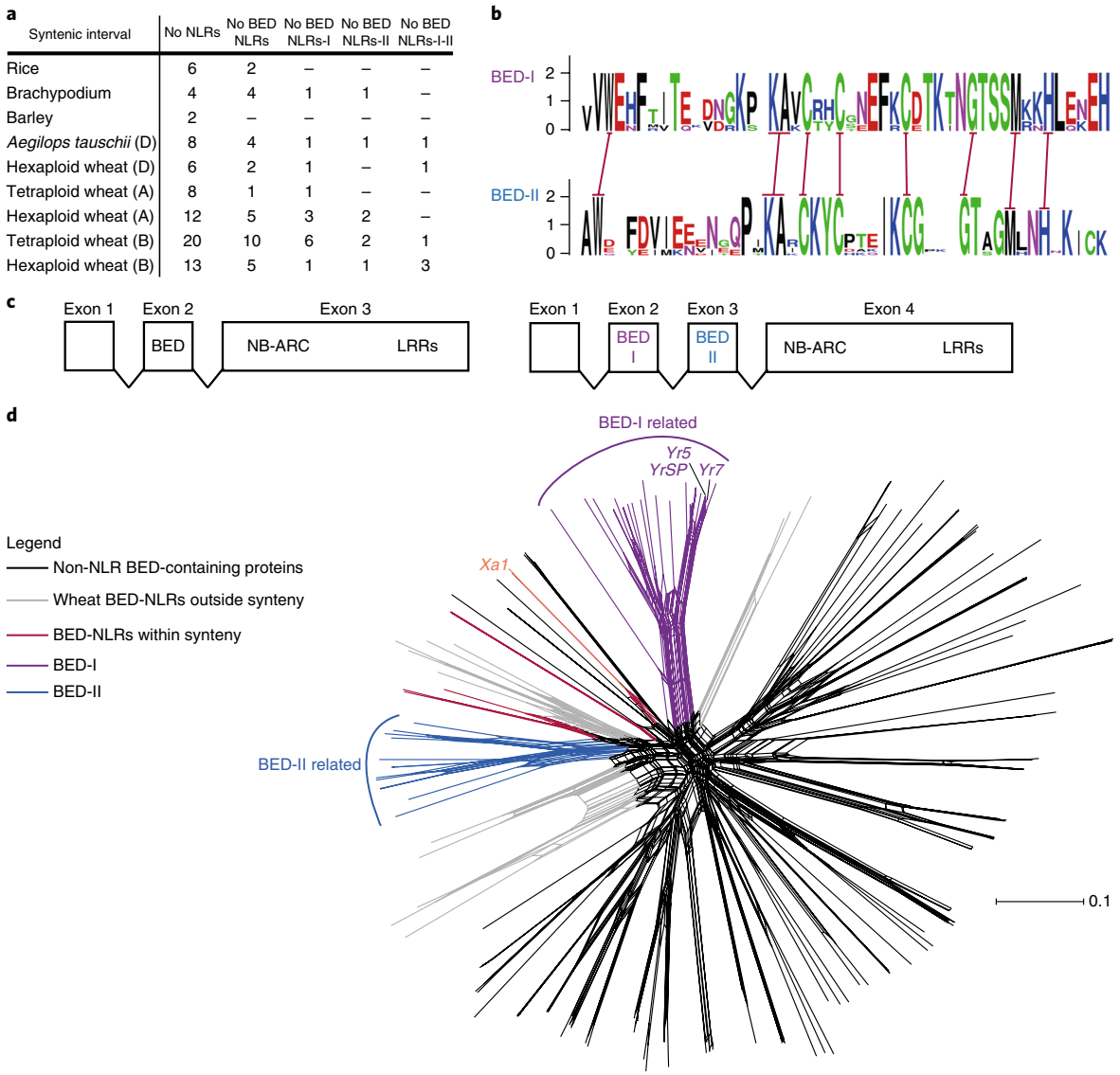


Fig. 3 | BED domains from BED-NLRs and non-NLR proteins are distinct a, Numbers of NLRs in the syntenic regions across grass genomes (see Supplementary Fig. 7), including BED-NLRs. **b**, WebLogo (http://weblogo.berkeley.edu/logo.cgi) diagram showing that the BED-I and BED-II domains are distinct, with only the highly conserved residues that define the BED domain (red bars) being conserved between the two types. **c**, Gene structure most commonly observed for BED-NLRs and BED-BED-NLRs within the *Yr7/Yr5/YrSP* syntenic interval. **d**, Neighbour-net analysis based on uncorrected P distances obtained from alignment of 153 BED domains including the 108 BED-containing proteins (including 25 NLRs) from RefSeq v1.0, BED domains from NLRs located in the syntenic region as defined in Supplementary Fig. 7, and BED domains from Xa1 and ZBED from rice. BED-I and II clades are highlighted in purple and blue, respectively. BED domains from the syntenic regions not related to either of these types are in red. BED domains derived from non-NLR proteins are in black and BED domains from BED-NLRs outside the syntenic region are in grey. Seven BED domains from non-NLR proteins were close to BED domains from BED-NLRs. Supplementary Fig. 9 includes individual labels.

We defined the *Yr7/Yr5/YrSP* syntenic interval across the wheat genomes and related grass species *Aegilops tauschii* (D genome progenitor), *Hordeum vulgare* (barley), *Brachypodium distachyon*, and *Oryza sativa* (rice) (Supplementary Files 4 and 5 and Supplementary Fig. 7). We identified both canonical NLRs and BED-NLRs across all genomes and species, except for barley, which contained canonical NLRs only across the syntenic region. The phylogenetic relationship based on the NB-ARC domain suggests a common evolutionary origin of these integrated domain NLR proteins before the wheat–rice divergence (~50 Mya) and an expansion in the number of NLRs in the A and B genomes of polyploid wheat species (Fig. 3a and Supplementary Fig. 8). Within the interval we also identified several genes in the A, B and D genomes that encode two consecutive in-frame BED domains (named BED-I and BED_II; Fig. 3b,c and Supplementary Fig. 7) followed by the canonical NLR. The BED domains in these genes were fully encoded within a single exon (exons 2 and 3) and in most cases had a four-exon structure (Fig. 3c). This is consistent with the three-exon structure of single BED-domain genes, such as *Yr7* and *Yr5/YrSP* (BED-I encoded on exon 2). This means we are able to report the double BED-domain NLR protein structure. The biological function of this molecular innovation remains to be determined, although our data show that the single BED-I structure can confer *Pst* resistance and is required for *Yr7*-mediated resistance.

Among other mechanisms, integrated domains of NLRs are hypothesized to act as decoys for pathogen effector targets[5]. This suggests that the integrated domain might be sequence-related to the host protein targeted by the effector. To identify these

potential effector targets in the host, we retrieved all BED-domain proteins (108) from the hexaploid wheat genome, including 25 BED-NLRs, and additional BED-NLRs located in the syntenic intervals (Supplementary Table 12 and Supplementary File 4). We also retrieved the rice Xa1[10,11] and ZBED proteins, the latter being hypothesized to mediate rice resistance to *Magnaporthe oryzae*[7]. We used the split network method implemented in SplitsTree4[30] to represent the relationships between these BED domains (Fig. 3d and Supplementary Figure 9). Overall, BED domains are diverse, although there is evidence of a split between BED domains from BED-NLRs and non-NLR proteins (only 7 of 83 non-NLRs clustered with the BED-NLRs). Given that the base of the split is broad, integrated BED domains are most probably derived from multiple integration events. However, *Yr7* and *Yr5/YrSP* both arose from a common integration event that occurred before the *Brachypodium*-wheat divergence (Supplementary Fig. 9, purple). This is consistent with the hypothesis that integrated domains might have evolved to strengthen the interaction with pathogen effectors after integration[31], although we cannot exclude the potential role of the BED domains in signalling at this stage.

Among BED-NLRs, BED-I and BED-II constitute two major clades, consistent with their relatively low amino-acid conservation (Fig. 3b), that comprise solely genes from within the *Yr7/Yr5/YrSP* syntenic region. Seven non-NLR BED-domain wheat proteins clustered with BED-NLRs. These are most closely related to the *Brachypodium* and rice BED-NLR proteins and were not expressed in RNA-Seq data from a *Yr5* time-course (re-analysis of published data[32]; Supplementary Fig. 10; and Supplementary Table 13). Similarly, no BED-containing protein was differentially expressed during this infection time-course, consistent with the prediction that effectors alter their targets' activity at the protein level in the integrated-decoy model[5]. We cannot however disprove that these closely related BED-containing proteins are involved in BED-NLR-mediated resistance.

BED-NLRs are frequent in Triticeae, and occur in other monocot and dicot tribes[6–8]. To date, a single BED-NLR gene, *Xa1*, has been shown to confer resistance to plant pathogens[10,11]. In the present study, we show that the distinct *Yr7*, *Yr5* and *YrSP* resistance specificities belong to a complex NLR cluster on chromosome 2B and are encoded by BED-NLRs genes that are linked in repulsion. We report five haplotypes for *Yr5/YrSP*, including three full-length BED-NLRs (including *Yr5*) and two truncated versions (including *YrSP*). These alternative haplotypes could be of functional significance as previously shown for the *Mla* and *Pm3* loci that confer resistance to *Blumeria graminis*[25,33] in barley and wheat, respectively, and the flax *L* locus conferring resistance to *Melampsora lini*[22]. Overall, our results add strong evidence for the importance of the BED-NLR architecture in plant–pathogen interactions. The relationship of these three distinct *Yr* loci will inform future hypothesis-driven engineering of novel recognition specificities.

Methods

MutRenSeq. *Mutant identification.* Supplementary Table 3 summarizes plant materials and *Pst* isolates used to identify mutants for each *Yr* gene. We used an EMS-mutagenized population in cultivar Cadenza[34] to identify mutants in *Yr7* using a forward genetic screen; whereas EMS-populations in the corresponding AvocetS-*Yr* near isogenic lines (NIL) were used to identify *Yr5* and *YrSP* mutants. For *Yr7*, we inoculated M_3 plants from the Cadenza EMS population with *Pst* isolate 08/21, which is virulent to *Yr1, Yr2, Yr3, Yr4, Yr6, Yr9, Yr17, Yr32, YrRob* and *YrSol*[35]. We hypothesized that susceptible mutants would carry mutations in *Yr7*. Plants were grown in 192-well trays in a confined glasshouse with no supplementary lights or heat. Inoculations were performed at the one leaf stage (Zadoks 11) with a talc-urediniospore mixture. Trays were kept in darkness at 10 °C and 100% humidity for 24 h. Infection types (IT) were recorded 21 days post-inoculation (dpi) following the Grassner and Straib scale[36]. Identified susceptible lines were progeny tested (12 to 16 plants per line) to confirm the reliability of the phenotype. DNA from all seven confirmed M_4 plants were used for RenSeq (see section below). Similar methods were used for AvocetS-*Yr7*, AvocetS-*Yr5* and AvocetS-*YrSP* EMS-mutagenized populations with the following exceptions: *Pst* pathotypes 108 E141A + (University of Sydney Plant Breeding Institute Culture no. 420), 150 E16A + (Culture no. 598) and 134 E16A + (Culture no. 572) were used to evaluate *Yr7*, *Yr5* and *YrSP* mutants, respectively. The seven EMS-derived susceptible mutants in Lemhi-*Yr5* were previously identified[37] and progeny tested. DNA from M_5 plants from all seven mutants was used for RenSeq.

DNA preparation, resistance gene enrichment and sequencing (RenSeq). We extracted total genomic DNA from young leaf tissue using the large-scale DNA extraction protocol from the McCouch Lab (https://ricelab.plbr.cornell.edu/dna_extraction) and a previously described method[38]. We checked DNA quality and quantity on a 0.8% agarose gel and with a NanoDrop spectrophotometer (Thermo Scientific). Arbor Biosciences performed the targeted enrichment of NLRs according to the MYbaits protocol using an improved version of the previously published Triticeae bait library available at github.com/steuernb/MutantHunter. Library construction was performed using the TruSeq RNA protocol v2 (Illumina 15026495). Libraries were pooled with one pool of samples for Cadenza mutants and one pool of eight samples for the Lemhi-*Yr5* parent and Lemhi-*Yr5* mutants. AvocetS-*Yr5* and AvocetS-*YrSP* wild-type, together with their respective mutants, were also processed according to the MYbaits protocol and the same bait library was used. All enriched libraries were sequenced on a HiSeq 2500 (Illumina) in High Output mode using 250 bp paired end reads and SBS chemistry. For the Cadenza wild-type, we generated data on an Illumina MiSeq instrument. In addition to the mutants, we also generated RenSeq data for Kronos and Paragon to assess the presence of *Yr5* in Kronos and *Yr7* in Paragon. Details of all the lines sequenced, alongside NCBI accession numbers, are presented in Supplementary Tables 4 and 14.

MutantHunter pipeline. We adapted the pipeline from https://github.com/steuernb/MutantHunter/ to identify candidate contigs for the targeted *Yr* genes. First, we trimmed the RenSeq-derived reads with trimmomatic[39] using the following parameters: ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 MINLEN:50 (v0.33). We made de novo assemblies of wild-type plant trimmed reads with the CLC assembly cell and default parameters apart from the word size (-w) parameter that we set to 64 (v5.0, http://www.clcbio.com/products/clc-assembly-cell/) (Supplementary Table 15). We then followed the MutantHunter pipeline detailed at https://github.com/steuernb/MutantHunter/. For Cadenza mutants, we used the following MutantHunter program parameters to identify candidate contigs: *-c* 20 *-n* 6 *-z* 1000. These options require a minimum coverage of 20x for SNPs to be called; at least six susceptible mutants must have a mutation in the same contig to report it as a candidate; small deletions were filtered out by setting the number of coherent positions with zero coverage to call a deletion mutant at 1000. The *-n* parameter was modified accordingly in subsequent runs with the Lemhi-*Yr5* datasets (*-n* 6).

To identify *Yr5* and *YrSP* contigs from Avocet mutants, we followed the MutantHunter pipeline with all default parameters, except in the use of CLC Genomics Workbench (v10) for reads QC, trimming, de novo assembly of Avocet wild-type and mapping all the reads against de novo wild-type assembly. Default MutantHunter parameters were used except that *–z* was set as 100. The parameter *–n* was set to 2 in the first run and then to 3 in the second run. Two *Yr5* mutants were most probably sibling lines as they carried identical mutations at the same position (Supplementary Fig. 2 and Supplementary Table 4).

For *Yr7* we identified a single contig with six mutations, however we did not identify mutations in line Cad0903. On examination of the *Yr7* candidate contig we predicted that the 5' region was likely to be missing (Supplementary Fig. 2). We thus annotated potential NLRs in the Cadenza genome assembly available from the Earlham Institute (Supplementary Table 6, http://opendata.earlham.ac.uk/Triticum_aestivum/EI/v1.1) with the NLR-Annotator program using default parameters (https://github.com/steuernb/NLR-Annotator). We identified an annotated NLR in the Cadenza genome with 100% sequence identity to the *Yr7* candidate contig, which extended beyond our de novo assembled sequence. We therefore replaced the previous candidate contig with the extended Cadenza sequence (100% sequence identity) and mapped the RenSeq reads from Cadenza wild-type and mutants as described above. This confirmed the candidate contig for *Yr7* as we retrieved the missing 5' region including the BED domain. The improved contig now also contained a mutation in the outstanding mutant line Cad0903 (Supplementary Fig. 2). The Triticeae bait library does not include integrated domains in its design so they are prone to be missed, especially when located at the ends of an NLR. Sequencing technology could also have accounted for this: MiSeq was used for Cadenza wild-type whereas HiSeq was chosen for Lemhi-*Yr5* and we recovered the 5' region in the latter, although coverage was lower than for the regions encoding canonical domains. In summary, we sequenced nine, ten and four mutants for *Yr7*, *Yr5* and *YrSP*, respectively, and identified for each target gene a single contig that accounted for all progeny-tested susceptible mutants.

Candidate contig confirmation and gene annotation. We sequenced the *Yr7, Yr5* and *YrSP* candidate contigs from the mutant lines (annotated in Supplementary Files 1 and 2) to confirm the EMS-derived mutations using primers documented in Supplementary Table 16. We first PCR-amplified the complete locus from the same DNA preparations as the ones submitted for RenSeq with the Phusion®

High-Fidelity DNA Polymerase (New England Biolabs) following the supplier's protocol (https://www.neb.com/protocols/0001/01/01/pcr-protocol-m0530). We then carried out nested PCR on the obtained product to generate overlapping 600–1,000 bp amplicons that were purified using the MiniElute kit (Qiagen). The purified PCR products were sequenced by GATC following the LightRun protocol (https://www.gatc-biotech.com/shop/en/lightrun-tube-barcode.html). Resulting sequences were aligned to the wild-type contig using ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/). This allowed us to curate the *Yr7* locus in the Cadenza assembly that contained two sets of unknown ('N') bases in its sequence, corresponding to a 39 bp insertion and a 129 bp deletion (Supplementary File 3), and to confirm the presence of the mutations in each mutant line.

We used HISAT2[40] (v2.1) to map RNA-Seq reads available from Cadenza and AvocetS-*Yr5*[32] to the RenSeq de novo assemblies with curated loci to define the structure of the genes. We used the following parameters: --no-mixed --no-discordant to map reads in pairs only. We used the --novel-splicesite-outfile to predict splicing sites that we manually scrutinized with the genome visualization tool IGV[41] (v2.3.79). Predicted coding sequences (CDS) were translated using the ExPASy online tool (https://web.expasy.org/translate/). This allowed us to predict the effect of the mutations on each candidate transcript (Fig. 1a and Supplementary Table 4). The long-range primers for both *Yr7* and *Yr5* loci were then used on the corresponding susceptible Avocet NIL mutants to determine whether the genes were present and carried mutations in that background (Fig. 1a and Supplementary Files 1 and 2).

Coiled coil domain prediction. To determine whether *Yr7*, *Yr5* and *YrSP* encode Coiled Coil (CC) domains we used the NCOILS prediction program[42] (v1.0, https://embnet.vital-it.ch/software/COILS_form.html) with the following parameters: MTIK matrix with applying a 2.5-fold weighting of positions a,d. We compared the profiles to those obtained with already characterized CC-NLR encoding genes *Sr33*[24], *Mla10*[23], *Pm3*[25] and *RPS5* [43](Supplementary Fig. 4). We also ran the program on *Yr7* and *Yr5* protein sequences where the BED domain was manually removed to determine whether its integration could have disrupted an existing CC domain. To further investigate whether *Yr7*, *Yr5* and *YrSP* encode CC domains we performed a BLASTP analysis[44] with their N-terminal region, from the methionine to the first amino acid encoding the NB-ARC domain, with or without the BED domain (Supplementary Fig. 4).

Genetic linkage. We generated a set of F_2 populations to genetically map the candidate contigs (Supplementary Table 3). For *Yr7* we developed an F_2 population based on a cross between the susceptible mutant line Cad0127 to the Cadenza wild-type (population size 139 individuals). For *Yr5* and *YrSP* we developed F_2 populations between AvocetS and the NILs carrying the corresponding *Yr* gene (94 individuals for *YrSP* and 376 for *Yr5*). We extracted DNA from leaf tissue at the seedling stage (Zadoks 11) following a previously published protocol[45] and Kompetitive Allele Specific PCR (KASP) assays were carried out as described[46]. The R/qtl package[47] was used to produce the genetic map based on a general likelihood ratio test and genetic distances were calculated from recombination frequencies (v1.41-6).

We used previously published markers linked to *Yr7*, *Yr5* and *YrSP* (WMS526, WMS501 and WMC175, WMC332, respectively[15,19,20]) in addition to closely linked markers WMS120, WMS191 and WMC360 (based on the GrainGenes database https://wheat.pw.usda.gov/GG3/) to define the physical region on the Chinese Spring assembly RefSeq v1.0 (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies). Two different approaches were used for genetic mapping depending on the material. For *Yr7*, we used the public data[34] for Cad0127 (www.wheat-tilling.com) to identify nine mutations located within the *Yr7* physical interval based on BLAST analysis against RefSeq v1.0. We used KASP primers when available and manually designed additional ones including an assay targeting the Cad0127 mutation in the *Yr7* candidate contig (Supplementary Table 16). We genotyped the Cad0127 F_2 populations using these nine KASP assays and confirmed genetic linkage between the Cad0127 *Yr7* candidate mutation and the nine mutations across the physical interval (Supplementary Fig. 3).

For *Yr5* and *YrSP*, we first aligned the candidate contigs to the best BLAST hit in an AvocetS RenSeq de novo assembly. We then designed KASP primers targeting polymorphisms between these sequences and used them to genotype the corresponding F_2 population (Supplementary Table 16). For both candidate contigs we confirmed genetic linkage with the previously published genetic intervals for these *Yr* genes (Supplementary Fig. 3). Allelism tests between *Yr7*, *Yr5* and *YrSP* are described in the Supplementary Information.

***Yr7, Yr5* and *YrSP* gene-specific markers.** The development of gene-specific markers is described in the Supplementary Information.

In silico mining for *Yr7* and *Yr5*. We used the *Yr7* and *Yr5* sequences to retrieve the best BLAST hits in the *T. aestivum* and *T. turgdium* wheat genomes listed in Supplementary Table 6. The best *Yr5* hits shared between 93.6% and 99.3% sequence identity, which was comparable to what was observed for alleles derived from the wheat *Pm3* (>97% identity)[21] and flax *L* (>90% identity)[22] genes. *Yr7* was identified only in Paragon and Cadenza (Supplementary Table 7; See Supplementary File 3 for curation of the Paragon sequence).

Analysis of the *Yr7* and *Yr5/YrSP* cluster on RefSeq v1.0. *Definition of syntenic regions across grass genomes.* We used NLR-Annotator to identify putative NLR loci on RefSeq v1.0 chromosome 2B and identified the best BLAST hits to *Yr7* and *Yr5* on RefSeq v1.0. Additional BED-NLRs and canonical NLRs were annotated in close physical proximity to these best BLAST hits. Therefore, to better define the NLR cluster we selected ten non-NLR genes located both distal and proximal to the region, and identified orthologs in barley, *Brachypodium* and rice in *EnsemblPlants* (https://plants.ensembl.org/). We used different % ID cutoffs for each species (>92% for barley, >84% for *Brachypodium* and >76% for rice) and determined the syntenic region when at least three consecutive orthologues were found. A similar approach was conducted for *Triticum* ssp[48] and *Ae. tauschii*[49] (Supplementary File 4).

Definition of the NLR content of the syntenic region. We extracted the previously defined syntenic region from the grass genomes listed in Supplementary Table 6 and annotated NLR loci[50] with NLR-Annotator. We maintained previously defined gene models where possible, but also defined new gene models that were further analysed through a BLASTx analysis to confirm the NLR domains (Supplementary Files 4 and 5). The presence of BED domains in these NLRs was also confirmed by CD-Search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

Phylogenetic and neighbour network analyses. Methods for the phylogenetic analyses are described in the Supplementary Information.

Transcriptome analysis. Methods for the transcriptomic analyses are described in the Supplementary Information.

Received: 10 May 2018; Accepted: 31 July 2018;
Published online: 27 August 2018

References

1. Oerke, E. C. Crop losses to pests. *J. Agric. Sci.* **144**, 31–43 (2006).
2. Hubbard, A. et al. Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biol.* **16**, 23 (2015).
3. Aravind, L. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem. Sci.* **25**, 421–423 (2000).
4. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).
5. Kourelis, J. & van der Hoorn, R. A. L. Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *Plant Cell*. https://doi.org/10.1105/tpc.17.00579 (2018).
6. Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D. G. & Krasileva, K. V. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol.* **14**, 8 (2016).
7. Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X. & Morel, J.-B. Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol.* **210**, 618–626 (2016).
8. Bailey, P. C. et al. Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biol.* **19**, 23 (2018).
9. Bundock, P. & Hooykaas, P. An *Arabidopsis* hAT-like transposase is essential for plant development. *Nature* **436**, 282–284 (2005).
10. Yoshimura, S. et al. Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl Acad. Sci. USA.* **95**, 1663–1668 (1998).
11. Das, B., Sengupta, S., Prasad, M. & Ghose, T. Genetic diversity of the conserved motifs of six bacterial leaf blight resistance genes in a set of rice landraces. *BMC Genet.* **15**, 82 (2014).
12. Law, C. N. Genetic control of yellow rust resistance in *T. spelta* Album. *Plant Breed. Institute, Cambridge, Annu. Rep.* **1975**, 108–109 (1976).
13. Johnson, R. & Dyck, P. L. Resistance to yellow rust in *Triticum spelta* var. Album and bread wheat cultivars Thatcher and Lee. *Colloq. l'INRA* (1984).
14. Zhang, P., McIntosh, R. A., Hoxha, S. & Dong, C. M. Wheat stripe rust resistance genes *Yr5* and *Yr7* are allelic. *Theor. Appl. Genet.* **120**, 25–29 (2009).
15. Feng, J. Y. et al. Molecular mapping of *YrSP* and its relationship with other genes for stripe rust resistance in wheat chromosome 2BL. *Phytopathology* **105**, 1206–1213 (2015).
16. Wellings, C. R. & McIntosh, R. A. *Puccinia striiformis* f. sp. *tritici* in Australasia: pathogenic changes during the first 10 years. *Plant Pathol.* **39**, 316–325 (1990).
17. Zhan, G. et al. Virulence and molecular diversity of the *Puccinia striiformis* f. sp. *tritici* population in Xinjiang in relation to other regions of western China. *Plant Dis.* **100**, 99–107 (2016).
18. Steuernagel, B. et al. Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nat. Biotechnol.* **34**, 652–655 (2016).
19. Sun, Q., Wei, Y., Ni, Z., Xie, C. & Yang, T. Microsatellite marker for yellow rust resistance gene *Yr5* in wheat introgressed from spelt wheat. *Plant Breed.* **121**, 539–541 (2002).

20. Yao, Z. J. et al. The molecular tagging of the yellow rust resistance gene *Yr7* in wheat transferred from differential host Lee using microsatellite markers. *Sci. Agric. Sin.* **39**, 1146–1152 (2006).
21. Brunner, S. et al. Intragenic allele pyramiding combines different specificities of wheat *Pm3* resistance alleles. *Plant J.* **64**, 433–445 (2010).
22. Ellis, J. G., Lawrence, G. J., Luck, J. E. & Dodds, P. N. Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* **11**, 495–506 (1999).
23. Bai, S. et al. Structure-function analysis of barley NLR immune receptor MLA10 reveals its cell compartment specific activity in cell death and disease resistance. *PLoS Pathog.* **8**, e1002752 (2012).
24. Periyannan, S. et al. The gene *Sr33*, an ortholog of barley *Mla* genes, encodes resistance to wheat stem rust race Ug99. *Science* **341**, 786–788 (2013).
25. Srichumpa, P., Brunner, S., Keller, B. & Yahiaoui, N. Allelic series of four powdery mildew resistance genes at the *Pm3* locus in hexaploid bread wheat. *Plant Physiol.* **139**, 885–895 (2005).
26. Sarris, P. F. et al. A plant immune receptor detects pathogen effectors that target WRKY transcription factors. *Cell* **161**, 1089–1100 (2015).
27. Wingen, L. U. et al. Establishing the A. E. Watkins landrace cultivar collection as a resource for systematic gene discovery in bread wheat. *Theor. Appl. Genet.* **127**, 1831–1842 (2014).
28. Reeves, J. C. et al. Changes over time in the genetic diversity of four major European crops - a report from the Gediflux Framework 5 project. In *Proc. 17th EUCARPIA Gen. Congr.* (Eds Grausgruber, J. V. H. & Ruckenbauer, P.) 3–7 (BOKU, 2004).
29. Ellis, J. G., Lagudah, E. S., Spielmeyer, W. & Dodds, P. N. The past, present and future of breeding rust resistant wheat. *Front. Plant Sci.* **5**, 641 (2014).
30. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
31. Ellis, J. G. Integrated decoys and effector traps: how to catch a plant pathogen. *BMC Biol.* **14**, 13 (2016).
32. Dobon, A., Bunting, D. C. E., Cabrera-Quio, L. E., Uauy, C. & Saunders, D. G. O. The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression. *BMC Genomics* **17**, 380 (2016).
33. Seeholzer, S. et al. Diversity at the *Mla* powdery mildew resistance locus from cultivated barley reveals sites of positive selection. *Mol. Plant-Microbe Interact.* **23**, 497–509 (2010).
34. Krasileva, K. V. et al. Uncovering hidden variation in polyploid wheat. *Proc. Natl Acad. Sci. USA.* **6**, E913–E921 (2017).
35. Hubbard, A. J., Fanstone, V. & Bayles, R. A. *UKCPVS 2009 Annual report* (NIAB, 2009).
36. Gassner, G. & Straib, W. *Die Bestimmung der biologischen Rassen des Weizengelbrostes* (Pucciniaglumarum f.sp. tritici *Schmidt Erikss. u. Henn*). *Arb. Biol. Reichsanst. Land: Forstwirtsch.* **20**, 141–163 (1932).
37. McGrann, G. R. D. et al. Genomic and genetic analysis of the wheat race-specific yellow rust resistance gene *Yr5*. *J. Plant Sci. Mol. Breed.* **3**, (2014).
38. Lagudah, E. S., Appels, R., Brown, A. H. D. & McNeil, D. The molecular–genetic analysis of *Triticum tauschii*, the D-genome donor to hexaploid wheat. *Genome* **34**, 375–386 (1991).
39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
40. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
41. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
42. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
43. Warren, R. F., Henk, A., Mowery, P., Holub, E. & Innes, R. W. A mutation within the leucine-rich repeat domain of the arabidopsis disease resistance gene RPS5 partially suppresses multiple bacterial and downy mildew resistance genes. *Plant Cell* **10**, 1439–1452 (1998).
44. Altschul, S. F. et al. Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **272**, 5101–5109 (2005).
45. Pallotta, M. A. et al. Marker assisted wheat breeding in the southern region of Australia. in *Proc. 10th Int. Wheat Genet. Symp. Instituto Sperimentale Cerealcoltura* (Eds Pogna, N. & McIntosh, R.A.) 789–791 (Istituto Sperimentale per la Cerealicoltura, 2003).
46. Ramirez-Gonzalez, R. H. et al. RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnol. J.* **13**, 613–624 (2015).
47. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).
48. Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
49. Luo, M.-C. et al. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498–502 (2017).
50. Jupe, F. et al. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics* **13**, 75 (2012).

Acknowledgements

This work was supported by the UK Biotechnology and Biological Sciences Research Council Designing Future Wheat programme BB/P016855/1 and the Grains Research and Development Corporation, Australia. C.M. was funded by a PhD studentship from Group Limagrain and J.Z. is funded by PhD scholarships from the National Science Foundation (NSF) and the Monsanto Beachell-Borlaug International Scholars Programs (MBBISP). We thank the International Wheat Genome Sequencing Consortium for allowing pre-publication access to the RefSeq v1.0 assembly and gene annotation. We thank J. Dubcovsky and X. Zhang (University of California, Davis) for providing *Yr5* cultivars. We thank the John Innes Centre Horticultural Services and Limagrain Rothwell staff for management of the wheat populations. We also thank S. Specel (Limagrain; Clermont-Ferrand) and R. Goram (JIC) for their help in designing and running KASP assays, and S. Hoxha (The University of Sydney) for technical assistance. This research was supported by the NBI Computing Infrastructure for Science (CiS) group in Norwich, UK.

Author contributions

C.M. performed the experiments to clone *Yr7* and *Yr5* and the subsequent analyses of their loci and BED domains, designed the gene-specific markers, analysed the genotype data in the studied panels, and designed and made the figures. J.Z. performed the experiments to clone *YrSP*, confirm the *Yr7* and *Yr5* genes in AvocetS-*Yr7* and AvocetS-*Yr5* mutants, and identified the full length of *Yr5* and *YrSP* with their respective regulatory elements. C.M. and J.Z. developed the gene-specific markers. P.Z. and R.M. performed the EMS treatment, isolation, and confirmation of *Yr7, Yr5* and *YrSP* mutants in AvocetS NILs. P.F. performed the pathology work on the Cadenza *Yr7* mutants and the mapping populations. B.S. helped with the NLR -Annotator analysis and provided the bait library for target enrichment and sequencing of NLRs. N.M.A. provided DNA samples for allelic variation studies. L.B. provided Lemhi-*Yr5* mutants. R.M., E.L., P.Z., B.W., S.B. and C.U. conceived, designed and supervised the research. C.M. and C.U. wrote the manuscript. J.Z., P.Z., R.M., B.W., N.M.A., L.B. and E.L. provided edits.

Competing interests

A patent application based on this work has been filed (United Kingdom Patent Application No. 1805865.1).

Data availability

The data that support the findings of this study are presented in the supplementary information. All sequencing data have been deposited in the NCBI Short Reads Archive under accession numbers listed in Supplementary Table 14 (SRP139043). Cadenza (*Yr7*) and Lemhi (*Yr5*) mutants are available through the JIC Germplasm Resource Unit (www.seedstor.ac.uk).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0236-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.U.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Cristobal Uauy

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	MutantHunter: pipeline to identify NLR-type resistance genes using RenSeq and EMS mutagenesis screens (https://github.com/steuernb/MutantHunter). Trimmomatic v0.33: trimmer for illumina sequence data. CLC assembly cell v5.0: read mapper to a reference and de novo assembly of next-generation sequencing data. NLR-Annotator: tool to annotate loci associated with NLRs in large sequences (https://github.com/steuernb/NLR-Annotator). ClustalOmega: multiple sequence alignment program (https://www.ebi.ac.uk/Tools/msa/clustalo/). ExPASy Translate tool: allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence (https://web.expasy.org/translate/). NCOILS v1.0: Coil coiled prediction program (https://embnet.vital-it.ch/software/COILS_form.html). blastn (v2.2.30+): Nucleotide-Nucleotide BLAST. CD-Search: conserved domain prediction program (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). MUSCLE (v3.8.31): multiple sequence alignment program. Jalview (v2.10.1): multiple sequence alignment visualistion program. Gblocks (v0.91b): eliminates poorly aligned positions and divergent regions of an alignment of DNA or protein sequences. RAxML (v8.2.10): program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees. Dendroscope (v3.5.9): viewer for rooted phylogenetic trees and networks. SplitsTree4 (v4.16): program for computing unrooted phylogenetic networks from molecular sequence data.

hmmer (v3.1b2): program for searching sequence databases for sequence homologs, and for making sequence alignments.
DESeq2 (v1.18.1): Differential gene expression analysis based on the negative binomial distribution.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data has been deposited in the NCBI Short Reads Archive under accession numbers listed in Supplementary Table 13 (SRP139043). Cadenza (Yr7) and Lemhi (Yr5) mutants are available through the JIC Germplasm Resource Unit (www.seedstor.ac.uk).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical analyses were performed on an already published datasets. All the information regarding sample size are available in the corresponding study (Dobon et al., 2016).
Data exclusions	No data were excluded from the analyses.
Replication	Mutant plants were confirmed by at least three rounds of independent phenotyping across three generations
Randomization	When phenotyping plants, we randomized mutant and wild-type controls as is normal practice.
Blinding	Blinding was performed when phenotyping the mutant lines for their resistance to Pst.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	All biological materials will be made available through deposition in independent seed repositories.