DISSERTATION

CHARACTERIZING HOST GENETIC RESISTANCE TO *WHEAT STREAK MOSAIC VIRUS* (WSMV) AND *FUSARIUM* WILT DISEASE

Submitted by

Yucong Xie

Department of Soil and Crop Sciences

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2021

Doctoral Committee:

Advisor: Stephen Pearce

Cristiana Argueso
María Muñoz-Amatriaín
Punya Nachappa

ABSTRACT

CHARACTERIZING HOST GENETIC RESISTANCE TO *WHEAT STREAK MOSAIC VIRUS* (WSMV) AND *FUSARIUM* WILT DISEASE

Crop production is limited by a variety of biotic stresses caused by pathogens. This study focuses on wheat streak mosaic disease in wheat, caused by the viral pathogen *Wheat streak mosaic virus* (WSMV), and *Fusarium* wilt disease in banana, caused by the fungal pathogen *Fusarium oxysporium* f.sp. *cubense* (*Foc*). In this dissertation, I applied genomic and transcriptomic tools to study the *Wsm2* locus that confers genetic resistance to WSMV. Analyzing exome and transcriptome reads from wheat lines carrying *Wsm2*, I characterized structural variations and identified unique transcripts specific to these *Wsm2* carrying lines. Moreover, examination of candidate genes within the *Wsm2* interval identified several tandemly duplicated candidate genes annotated as Bowman-Birk inhibitor (BBIs), which triggered my interests to perform a genome-wide characterization of this gene family in wheat. I studied the possible mechanisms behind its copy number and functional domain duplications and analyzed its diverse role in plant biotic and abiotic stress using wheat RNA-seq expression data. Finally, I analyzed a time course transcriptomic dataset from banana root infected with *Foc* subtropical race 4 strain (*Foc*-STR4). I used gene co-expression assembly network (WGCNA) to study host plant transcriptional response to *Foc* infection and analyzed the expression profiles of candidate genes underlying a novel locus conferring resistance to *Foc*-STR4 and prioritized candidates. In summary, this dissertation studied genetic variants underlying host genetic resistance to WSMV and *Foc* and shed light on plant defense mechanisms against these two important crop pathogens.

AKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Dr. Stephen Pearce, for providing guidance and numerous supports throughout my PhD career. Thank you for always challenging me to think critically and encouraging me to become an independent scientist.

I would like to thank my committee, Dr. Cristiana Argueso, who led me into the Molecular Plant Microbe Interaction field, and Dr. Punya Nachappa, who taught me Plant Insect Vector Interaction. Dr. María Muñoz-Amatriaín, it's a great pleasure to work with you as a teaching assistant in the Plant Genetics course. I appreciate all of you for devoting your previous time and effort to my professional development.

I would like to thank the former Pearce Lab members, Dr. Rocío Alarcón-Reverte, Andrew Katz, Carl VanGessel, Caitlynd Krosch, Forrest Wold-McGimsey, and especially to Dr. Karl Ravet, who provided me with professional guidance during the early year of my PhD.

I also like to share my gratitude to Dr. Tessa Albrecht and Dr. Alyx Shigenaga who taught me pathogen inoculation assays. I would like to thank Dr. Scott Haley for sharing valuable plant materials for my research. I would like to thank Dr. Robyn Roberts and Dr. Yuan Zeng who provided me guidance for choosing my career path. I would like to thank my collaborators, Dr. Elizabeth Aitken and Dr. Shiwei Chen. Besides, I would also thank Courtland Kelly who helped and encouraged me in the last semester of my PhD study.

Last but not least, I am grateful for my friends and family, whose unconditional companionship, near or far, has supported me throughout the pandemics and encouraged me towards arriving at the milestone of my PhD study.

TABLE OF CONTENTS

ABSTRACT ii

AKNOWLEDGEMENTS iii

CHAPTER1. INTRODUCTION 1
1.1 Background 1
1.2 Molecular Basis of Plant Biotic Stress Resistance 3
1.3 Natural Genetic Variation is Important Source of Genetic Resistance 4
1.4 Dissertation Overview 8

REFERENCES 10

CHAPTER 2. EXTENSIVE STRUCTURAL VARIATION IN THE BOWMAN-BIRK INHIBITOR FAMILY IN COMMON WHEAT (*TRITICUM AESTIVUM* L.) 20
2.1 Summary 20
2.2 Introduction 21
2.3 Results 25
2.4 Discussions and Conclusion 36
2.5 Methods 42

CHAPTER 2 FIGURES 48

REFERENCES 56

CHAPTER 3. GENOMIC AND MOLECULAR CHARACTERIZATION OF THE *WHEAT STREAK MOSAIC VIRUS* RESISTANCE LOCUS 2 (*WSM2*) IN COMMON WHEAT (*TRITICUM AESTIVUM*. L) 72
3.1 Summary 72
3.2 Introduction 73
3.3 Materials and Methods 78
3.4 Results 85
3.5 Discussion 101

REFERENCES 106

CHAPTER 4. TRANSCRIPTOMICS OF BANANA (*MUSA ACCUMINATA*) IN RESPONSE TO *FUSARIUM OXYSPORUM* F.SP. *CUBENSE* (*FOC*) SUBTROPICAL RACE 4 (STR4) INFECTION 122
4.1 Summary 122
4.2 Introduction 123
4.3 Materials and Methods 125
4.4 Results 127
4.5 Discussion 135

REFERENCES 139

APPENDIX A Supplementary Materials for Chapter 2 145
Appendix A.1 - Supplementary figures 145
Appendix A.2 – Supplementary tables for chapter 2 (.xls) 148

APPENDIX B Supplementary Materials for Chapter 3 150
Appendix B.1 - Supplementary figures 150
Appendix B.2 – Supplementary tables for chapter 3 (.xls) 154

APPENDIX C Supplementary Materials for Chapter 4 155
Appendix C.1 - The R script used to run WGCNA analysis (wgcna_standard.R). 155
Appendix C.2 – Supplementary tables for chapter 4 (.xls) 155

CHAPTER 1. INTRODUCTION

Crops suffer from a variety of diseases that affect both yield and quality. During my PhD, I studied host genetic resistance to *Wheat streak mosaic virus* (WSMV) in wheat and *Fusarium* wilt disease in banana which is caused by fungal pathogen *Fusarium oxysporium* f.sp. *cubense* (*Foc*). This chapter serves as an introduction to disease background, plant immunity mechanisms, and current knowledge on wheat and banana genomic resources. In my research, I applied genetic, genomic and bioinformatic tools to characterize genetic variation underlying QTLs that confer resistance to WSMV in wheat and *Foc* in banana and used transcriptomic approaches to explore plant resistance mechanisms. Results from these projects are described in Chapters 2 – 4.

1.1 Background

There is an urgent need to sustainably increase global food production and meet the challenge to feed over 9 billion people by 2050 (J et al., 2010). Bread wheat (*Triticum aestivum* L.) is one of the three major cereals that dominate global food production, providing approximately 20% of all calories consumed as well as significant amounts of protein and nutrients for human consumption (FAOSTAT, 2020). Banana (*Musa* sp.) is the leading fruit crop in world agricultural production and trade and is an important staple food source in developing countries (FAOSTAT, 2020).

Among the most important factors limiting crop production are biotic stresses that can cause plant disease, including viruses, fungi, bacteria, nematodes, and insects (Singla & Krattinger, 2015). Fungal pathogens are a major source of disease to wheat production and cause leaf rust (*Puccinia triticina* Eriks), stem rust (*P. graminis* Pers. f. sp. *tritici*), stripe rust (*P. striiformis* Westend f. sp. *tritici*), and powdery mildew (*Blumeria graminis* f. sp. *tritici*) (Narang et al., 2020; Singla & Krattinger, 2015). Viral diseases also threaten worldwide wheat production and include the yellow dwarf disease (YDD) that is caused mainly by *Barley Yellow Dwarf Virus* (*BYDV*,

genus, *Luteovirus*, family, *Luteoviridae*) (Miller and & Rasochová, 1997) and wheat streak mosaic disease (WSMD) that is caused by infection with *Wheat Streak Mosaic Virus* (WSMV, genus, *Tritimovirus*, family, *Potyviridae*) (Singh & Kundu, 2018). The major constraint for global banana production is *Fusarium* wilt disease, also known as Panama disease, which is caused by the fungal pathogen *Fusarium oxysporium* f.sp. *cubense* (*Foc*) (Ploetz, 2006). Disease symptoms for WSMV and *Foc* are illustrated in Figure 1.1.

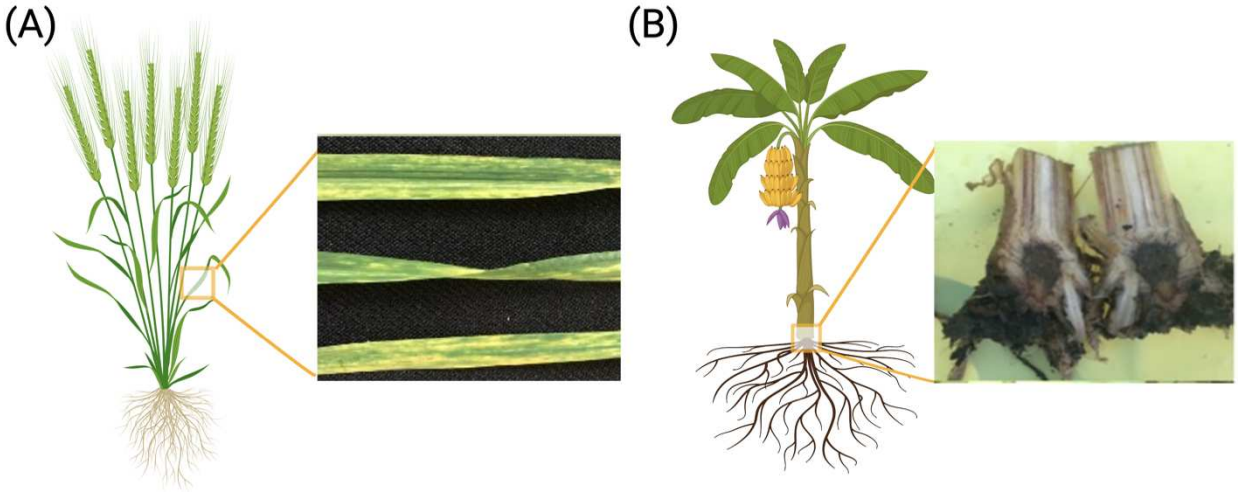


Figure 1.1. Disease symptoms of WSMV and *Foc* infection. **(A)** The viral pathogen WSMV affects wheat leaves, and infected plants show a yellowed mosaic pattern with discontinuous streaks. **(B)** The soil-borne fungal pathogen *Foc* affect banana root, and lead to leaf discoloration, yellowing and wilt along the edges of leaves.

Different disease management tools are available to control biotic stresses. Chemical control, such as using fungicides, miticides, and pesticides, is not always available for many diseases, such as Wheat streak mosaic disease and *Fusarium* wilt disease (Ploetz, 2015; Singh & Kundu, 2018), and are not environmentally friendly. Moreover, cultural practices, such as removing diseased individuals, can be costly and laborious. In comparison, genetic resistance is the most effective and sustainable long-term management option to reduce yield losses caused from various biotic stresses (Bailey-Serres et al., 2019). In wheat, although some disease resistance loci have been identified with molecular markers, only a small number of underlying resistance (*R*) genes have been cloned and characterized (Bakala et al., 2021). The evolution of genomic approaches has

accelerated cloning of *R* genes from wild relatives, using techniques such as *R* gene enrichment sequencing (RenSeq), Associated Genetics Renseq (AgRenSeq) (Arora et al., 2019) and Mutagenesis Renseq (MutRenSeq) (Armstrong et al., 2019). However, introgression of these genes into elite crop cultivars remains laborious with traditional breeding methods, and technically difficult with transgenic approaches as many wheat cultivars remain recalcitrant to tissue culture and regeneration (Gao, 2021). As crops and pathogens compete in their evolution, the pathogens may overcome the resistance provided by single *R* genes, making the *R* gene ineffective against newly evolved pathogen strains (Frank, 1992). Cloning and characterization of resistance genes helps us to understand plant immunity mechanisms. Greater access to more cloned *R* genes could be engineered into crops as a stack to help breed for strong, durable, and broad-spectrum resistant crops to fight against various pathogens.

1.2 Molecular Basis of Plant Biotic Stress Resistance

Plant innate immunity against pathogens can be categorized into two main groups. The first layer of resistance is triggered by perception of pathogen or microbial associated molecular patterns (PAMPs/MAMPs) through plant transmembrane pattern recognition receptors (PRRs) which results in PAMP-triggered immunity (PTI) (Zipfel, 2014; Couto & Zipfel, 2016; Sánchez-Martín & Keller, 2019). Pathogens can release effectors to suppress PTI, leading to effector triggered susceptibility (ETS) (Jones & Dangl, 2006). In turn, plants have evolved resistance (*R*) genes that can recognize pathogen avirulence effectors (Avr), leading to effector-triggered immunity (ETI) (Jones & Dangl, 2006).

Plant immune responses downstream of PTI and ETI overlap and include calcium ion influx, reactive oxygen species (ROS) production, mitogen-activated protein kinase (MAPK)-mediated signaling, transcriptional reprogramming and hormone accumulation (Bigeard et al., 2015; Cui et

al., 2015). Important hormone-mediated plant immune responses include the antagonistic crosstalk between salicylic acid (SA) and jasmonic acid (JA) pathways (Shigenaga et al., 2017). SA accumulation induced plant ETI-mediated resistance against biotrophic and hemi-biotrophic pathogens, whereas JA and ethylene (ET) biosynthesis and signaling usually induces PTI-mediated resistance against necrotrophic pathogens (Glazebrook, 2005).

Most dominant *R* genes in plants cloned to date encode nucleotide-binding site (NBS) and leucine-rich repeats (LRRs) proteins, which directly or indirectly interact with pathogenic Avr effectors to trigger downstream defense responses (Balconi et al., 2012). The NBS-LRR proteins (NLRs) contain three main domains: the central nucleotide binding domain (NB-ARC) is highly conserved and binds ADP or ATP molecules, LRRs at the C-terminus that are important for recognition specificity, and a highly diverse N-terminus with two main domain groups, the coiled-coil (CC) or Toll and Interleukin-1 Receptor (TIR) domains that are responsible for downstream signaling after pathogen recognition (Marone et al., 2013; de Ronde et al., 2014). *R* gene-mediated resistance often induces hypersensitive responses (HR) that includes programmed cell death (PCD) and causes detectable necrotic lesions in the infected tissues (Greenberg & Yao, 2004). In rare occasions of plant resistance to some viral pathogens, the *R* gene mediated ETI response does not trigger observable necrosis or HR, defined as extreme resistance (ER) (de Ronde et al., 2014).

1.3 Natural Genetic Variation is an Important Source of Genetic Resistance

The long-term strategy to develop crops with genetic resistance to biotic stress often relies on exploring natural genetic diversity for resistance alleles followed by introgression of such resistance genes into elite varieties by recombination. Genetic variation includes single nucleotide polymorphism (SNPs), small insertion and deletions (Indels), as well as larger structural variations (SVs) such as copy number variations (CNVs), gene presence/absence variations (PAVs), and

chromosomal rearrangements (Escaramís et al., 2015; Yuan et al., 2021). Advances in high-throughput sequencing technologies have facilitated the identification of SVs in plants, and there is growing evidence showing that SVs are important sources of genetic variation associated with disease resistance traits (Wellenreuther et al., 2019). Genome wide analysis of SVs have shown *R* genes are highly enriched for CNVs in many plant species, including soybean (E et al., 2012; McHale et al., 2012; Lee et al., 2015), barley (Muñoz-Amatriaín et al., 2013), maize (Richter et al., 1995; Beló et al., 2009), and rice (Yu et al., 2013). CNVs can be one of the possible mechanisms to enhance plant disease resistance through duplication or deletion of *R* genes to modify gene expression levels.

1.3.1 Accessing natural genetic variation in polyploid wheat

As one of the major cereal crops, diseases in wheat pose a particularly large challenge to maintaining food security. An expanding set of genomic resources facilitates the study of diverse sources of genetic variation that can be applied to develop wheat cultivars with genetic resistance to important pathogens. Two major classes of domesticated wheat account for the vast majority of production: tetraploid durum wheat (*Triticum turgidum* ssp.*durum*; AABB genome) used for pasta, and hexaploid common wheat (*Triticum aestivum* L.; AABBDD genome) used for making bread, noodles and cookies (Jorge & Jan, 2007). Common wheat originates from two separate hybridization events (Thomas et al., 2014). The first occurred approximately 0.5 to 0.9 million years ago between *Triticum urartu* (AA genome) and an unknown species related to *Aegilops speltoides* (BB genome) which gave rise to *Triticum turgidum* ssp. *dicoccoides* (AABB genome). The second hybridization event between *Triticum turgidum* ssp. *durum* (AABB) and *Aegilops tauschii* (DD) occurred approximately 10,000 years ago (El Baidouri et al., 2017).

Genetic studies in wheat species are complicated due to its large and polyploid genome (16 Gb for common wheat and 12 Gb for durum wheat) and highly repetitive DNA sequences (> 85%) (Appels et al., 2018; Maccaferri et al., 2019). Central to many genetic studies is the most complete and best annotated chromosome-level genome assembly of the common wheat landrace 'Chinese Spring' (IWGSC RefSeq v1.0), which has a total assembly size of 14.5 Gb, representing 94% of the whole genome and includes 107,891 high-confidence (HC) gene models and 161,537 low-confidence (LC) gene models (Appels et al., 2018). A recent optical map based on long-read sequencing refined the 'Chinese Spring' reference genome as IWGSC RefSeq v2.1 with 108,010 HC and 161,535 LC gene models on this assembly (Zhu et al., 2021).

In addition to 'Chinese Spring', sixteen other wheat lines have publicly available reference-quality genome assemblies, representing the wheat pangenome (Walkowiak et al., 2020). These resources facilitate the study of within-species genomic variation and haplotype analysis (Brinton et al., 2020). Moreover, the genome sequence of wild emmer wheat (*Triticum turgidum* ssp. *dicoccoides*, AABB genome) 'Zavitan' (Avni et al., 2017) and a modern durum wheat 'Svevo' (Maccaferri et al., 2019) has been released, together with the genome sequence of diploid ancestral progenitors *Triticum urartu* (AA genome, Ling et al., 2018) and *Aegilops tauschii* (DD genome, Luo et al., 2017), which are important genetic resources to study wheat evolution and to explore genetic diversity for rare disease resistance alleles in wild species.

To dissect natural genetic variation underlying important QTLs in wheat lines without genome assemblies, reduced representation methods such as genotyping-by-sequencing (GBS) is attractive and is routinely applied in wheat breeding programs instead of whole genome sequences (WGS) (Poland & Rife, 2012). GBS can reduce the genome sequencing complexity with relatively low cost but still be able to discover genetic polymorphisms that span whole genomes from specific

wheat cultivars (Poland & Rife, 2012). Another reduced representation method for genetic variant discovery in wheat is exome sequencing that targets the coding regions of the genome (Biesecker et al., 2011; Winfield et al., 2012; Gardiner et al., 2019). Exome capture sequencing can be used to detect SNP variants as well as SVs (Saintenac et al., 2011). The detected genome-wide genetic polymorphisms, especially SNPs, can be utilized for association mapping, i.e., genome-wide association studies (GWAS) (Huang & Han, 2014; Ogura & Busch, 2015) and refined linkage mapping, i.e., quantitative trait locus (QTL) mapping (Hussain et al., 2017). For example, linkage and association analysis using iSelect SNP array identified a locus on wheat chromosome 6D that provides resistance to wheat curl mite (Dhakal et al., 2018).

In addition to genomic approaches to discover genetic variation, transcriptomics is also a powerful approach to study candidate genes underlying the QTLs using RNA transcript levels (Lowe et al., 2017). RNA sequencing (RNA-seq), which uses high-throughput sequencing to quantify transcripts in biological samples, can be used to quantify gene expression during development and under different conditions, and can also be used to discover novel transcripts or splice variants in *de novo* transcriptome assemblies (Wang et al., 2009). Furthermore, RNA-seq data can be used to assemble gene co-expression networks to reveal the gene regulation processes and to predict function of uncharacterized genes (Borrill et al., 2015). The wheat gene expression atlas contains over 900 RNA-seq samples from various tissue types, spanning different developmental stages, including multiple biotic and abiotic stresses and cultivars (Borrill et al., 2016; Ramírez-González et al., 2018). This RNA-seq database facilitates the study of potential molecular functions of orthologous genes of interest and provides sources to narrow down candidate genes within a QTL region (Borrill et al., 2019).

1.3.2 Genomic resources for banana

Similar to wheat, the banana genome is also polyploid and most cultivated bananas have a triploid genome (2n = 3x = 33, genome constitutions of AAA, AAB, or ABB), derived from two diploid progenitors, *Musa acuminata* (AA genome) and *Musa balbisiana* (BB genome) (D'Hont et al., 2000). The first *Musa* reference genome is of DH-Pahang, a doubled haploid *Musa acuminata* genotype (2n = 22, AA genome), of the subspecies *malaccensis* with a genome of 523 Megabases (Mbp) containing 36,542 protein-coding gene models (D'hont et al., 2012). Other genomic resources and tools for banana include several other *Musa* genome assemblies, transcriptomics and metabolic pathways, and genetic variants compared to rice that are available on the Banana Genome Hub (Droc et al., 2013), which facilitates studies to understand the basis of disease resistance in banana.

1.4 Dissertation Overview

During my PhD, I researched the *Wsm2* locus that confers WSMV resistance in wheat. An examination of candidate genes within this locus revealed extensive duplication of Bowman-Birk inhibitor (BBI) genes, a family of plant protease inhibitors (PIs), which triggered my interest to characterize this gene family in wheat. I used genomic and transcriptomic tools to characterize genetic variants underlying *Wsm2*. I also worked on a collaborative project with Dr. Elizabeth Aitken's lab at University of Queensland in Australia to study banana *Fusarium* wilt disease resistance.

In Chapter 2, I performed a genome wide characterization of the BBI gene family in common wheat and used phylogenetics to compare with the orthologs in rice, maize, barley and *Brachypodium* (Xie et al., 2021). I explored SVs of this gene family in wheat progenitors and among wheat cultivars and identified extensive CNVs and PAVs as well as small genetic variation

that leads to tandem functional domain duplications. The expression profiles of BBIs in common wheat were explored using RNA-seq databases, showing that members of this gene family likely have diverse functions throughout wheat development, and in response to biotic and abiotic stress.

In Chapter 3, I performed genomics and transcriptomics characterization of the *Wsm2* locus. Haplotype analysis in wheat pangenomes revealed that the genomic region underlying the *Wsm2* locus is highly dynamic among wheat cultivars and the variants conferring WSMV resistance are likely to be rare. I analyzed exome capture reads from 'Snowmass' (*Wsm2*+) and identified genetic polymorphisms and CNVs for candidate genes underlying *Wsm2* locus. I performed an RNA-seq study and explored unmapped reads using a *de novo* transcriptome assembly of *Wsm2*+ lines to identify unique transcripts specific to *Wsm2*+ line. From the RNA-seq study, I also explored host responses to WSMV infection and identified five candidates within *Wsm2* interval. My study sheds light on possible causative genes underlying *Wsm2* and facilitates hypothesis generation of candidates that can be tested to breed WSMV resistant wheat cultivar. Moreover, this work also developed CRISPR/Cas9-edited gene knockout mutants to test one of the top selected causative gene, *RPM1*, underlying *Wsm2* locus.

In Chapter 4, I analyzed a time course transcriptome dataset for resistant and susceptible banana genotypes in response to *Foc*-STR4 infection and used gene co-expression network assembly (WGCNA) to study overall transcriptomic changes. I found that the reactive oxygen species (ROS) production and cell wall strengthening related responses were induced in both resistant and susceptible banana genotypes. Comparatively, the resistant genotype induces such defense responses much faster and transiently than in susceptible genotype. Moreover, the transcriptome analysis also helped to prioritize candidates underlying a novel *Foc*-STR4 resistance locus identified from a QTL-seq study.

REFERENCES

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A. A., Himmelbach, A., Balfourier, F., … Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 361(6403), 7191–7191. https://doi.org/10.1126/science.aar7191

Armstrong, M. R., Vossen, J., Lim, T. Y., Hutten, R. C. B., Xu, J., Strachan, S. M., Harrower, B., Champouret, N., Gilroy, E. M., & Hein, I. (2019). Tracking disease resistance deployment in potato breeding by enrichment sequencing. Plant Biotechnology Journal, 17(2), 540–549. https://doi.org/10.1111/pbi.12997

Arora, S., Steuernagel, B., Gaurav, K., Chandramohan, S., Long, Y., Matny, O., Johnson, R., Enk, J., Periyannan, S., Singh, N., Asyraf Md Hatta, M., Athiyannan, N., Cheema, J., Yu, G., Kangara, N., Ghosh, S., Szabo, L. J., Poland, J., Bariana, H., … Wulff, B. B. H. (2019). Resistance gene cloning from a wild crop relative by sequence capture and association genetics. Nature Biotechnology, 37(2), 139–143. https://doi.org/10.1038/s41587-018-0007-9

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., Jordan, K. W., Golan, G., Deek, J., Ben-Zvi, B., Ben-Zvi, G., Himmelbach, A., Maclachlan, R. P., Sharpe, A. G., Fritz, A., … Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science, 357(6346), 93–97. https://doi.org/10.1126/science.aan0032

Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D., & Schroeder, J. I. (2019). Genetic strategies for improving crop yields. Nature, 575(7781), 109–118. https://doi.org/10.1038/s41586-019-1679-0

Bakala, H. S., Mandahal, K. S., Ankita, Sarao, L. K., & Srivastava, P. (2021). Breeding wheat for biotic stress resistance: achievements, challenges and prospects. Current Trends in Wheat Research. https://doi.org/10.5772/intechopen.97359

Balconi, C., Stevanato, P., Motto, M., & Biancardi, E. (2012). Breeding for biotic stress resistance/tolerance in plants. Crop Production for Agricultural Improvement (pp. 57–114). Springer Netherlands. https://doi.org/10.1007/978-94-007-4116-4_4

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., & Rafalski, A. (2009). Allelic genome structural variations in maize detected by array comparative genome hybridization. Theoretical and Applied Genetics, 120(2), 355–355. https://doi.org/10.1007/s00122-009-1128-9

Biesecker, L. G., Shianna, K. V., & Mullikin, J. C. (2011). Exome sequencing: The expert view. Genome Biology, 12(9), 128–128. https://doi.org/10.1186/gb-2011-12-9-128

Bigeard, J., Colcombet, J., & Hirt, H. (2015). Signaling mechanisms in pattern-triggered immunity (PTI). Molecular Plant, 8(4), 521–539. https://doi.org/10.1016/j.molp.2014.12.022

Borrill, P., Adamski, N., & Uauy, C. (2015). Genomics as the key to unlocking the polyploid potential of wheat. New Phytologist, 208(4), 1008–1022. https://doi.org/10.1111/nph.13533

Borrill, P., Harrington, S. A., & Uauy, C. (2019). Applying the latest advances in genomics and phenomics for trait discovery in polyploid wheat (Vol. 97, Issue 1, p. 72). https://doi.org/10.1111/tpj.14150

Borrill, P., Ramirez-Gonzalez, R., & Uauy, C. (2016). ExpVIP: A customizable RNA-seq data analysis and visualization platform. Plant Physiology, 170(4), 2172–2186. https://doi.org/10.1104/pp.15.01667

Brinton, J., Ramirez-Gonzalez, R. H., Simmonds, J., Wingen, L., Orford, S., Griffiths, S., Haberer, G., Spannagl, M., Walkowiak, S., Pozniak, C., & Uauy, C. (2020). A haplotype-led approach to increase the precision of wheat breeding. Communications Biology, 3(1). https://doi.org/10.1038/s42003-020-01413-2

Couto, D., & Zipfel, C. (2016). Regulation of pattern recognition receptor signaling in plants. Nature Reviews Immunology, 16(9), 537–552. https://doi.org/10.1038/nri.2016.77

Cui, H., Tsuda, K., & Parker, J. E. (2015). Effector-triggered immunity: from pathogen perception to robust defense. Annual Review of Plant Biology, 66(1), 487–511. https://doi.org/10.1146/annurev-arplant-050213-040012

de Ronde, D., Butterbach, P., & Kormelink, R. (2014). Dominant resistance against plant viruses. Frontiers in Plant Science, 5, 1–17. https://doi.org/10.3389/fpls.2014.00307

Dhakal, S., Tan, C. T., Anderson, V., Yu, H., Fuentealba, M. P., Rudd, J. C., Haley, S. D., Xue, Q., Ibrahim, A. M. H., Garza, L., Devkota, R. N., & Liu, S. (2018). Mapping and KASP marker development for wheat curl mite resistance in "TAM 112" wheat using linkage and association analysis. Molecular Breeding, 38(10), 119–119. https://doi.org/10.1007/s11032-018-0879-x

D'hont, A., Denoeud, F., Aury, J. M., Baurens, F. C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., Da Silva, C., Jabbari, K., Cardi, C., Poulain, J., Souquet, M., Labadie, K., Jourda, C., Lengellé, J., Rodier-Goud, M., … Wincker, P. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature, 488(7410), 213–217. https://doi.org/10.1038/nature11241

D'Hont, A., Paget-Goy, A., Escoute, J., & Carreel, F. (2000). The interspecific genome structure of cultivated banana, *Musa* spp. Revealed by genomic DNA in situ hybridization. Theoretical and Applied Genetics, 100(2), 177–183. https://doi.org/10.1007/s001220050024

Droc, G., Larivière, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J.-F., Lengelle, J., Baurens, F.-C., Cenci, A., Pitollat, B., D'Hont, A., Ruiz, M., Rouard, M., & Bocs, S. (2013). The banana genome hub. database, 2013. https://doi.org/10.1093/database/bat035

E, C. D., Geon, L. T., Xiaoli, G., Sara, M., Kai, W., M, B. A., Jianping, W., J, H. T., K, W. D., E, C. T., W, D. B., Jiming, J., E, H. M., & F, B. A. (2012). Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. Science, 338(6111), 1206–1209. https://doi.org/10.1126/science.1228746

El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., Alaux, M., Quesneville, H., Pont, C., & Salse, J. (2017). Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). New Phytologist, 213(3), 1477–1486. https://doi.org/10.1111/nph.14113

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: description, history, and methods to detect structural variation. Briefings in Functional Genomics, 14(5), 305–314. https://doi.org/10.1093/bfgp/elv014

FAOSTAT. (2020). Food and agriculture organization of the united nations database, FAOSTAT statistics, Crops-FAOSTAT statistics, Crops. https://doi.org/10.1016/B978-0-12-384947-2.00270-1

Frank, S. A. (1992). Models of plant-pathogen coevolution. Trends in Genetics, 8(6), 213–219. https://doi.org/10.1016/0168-9525(92)90236-W

Gao, C. (2021). Genome engineering for crop improvement and future agriculture. Cell, 184(6), 1621–1635. https://doi.org/10.1016/j.cell.2021.01.005

Gardiner, L.-J., Brabbs, T., Akhunov, A., Jordan, K., Budak, H., Richmond, T., Singh, S., Catchpole, L., Akhunov, E., & Hall, A. (2019). Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. GigaScience, 8(4). https://doi.org/10.1093/gigascience/giz018

Glazebrook, J. (2005). Contrasting Mechanisms of defense against biotrophic and necrotrophic pathogens. Annual Review of Phytopathology, 43(1), 205–227. https://doi.org/10.1146/annurev.phyto.43.040204.135923

Greenberg, J. T., & Yao, N. (2004). The role and regulation of programmed cell death in plant–pathogen interactions. Cellular Microbiology, 6(3), 201–211. https://doi.org/10.1111/j.1462-5822.2004.00361.x

Huang, X., & Han, B. (2014). Natural variations and genome-wide association studies in crop plants. Annual Review of Plant Biology, 65(1), 531–551. https://doi.org/10.1146/annurev-arplant-050213-035715

Hussain, W., Stephen Baenziger, P., Belamkar, V., Guttieri, M. J., Venegas, J. P., Easterly, A., Sallam, A., & Poland, J. (2017). Genotyping-by-sequencing derived high-density linkage map and its application to QTL mapping of flag leaf traits in bread wheat. Scientific Reports, 7(1), 1–15. https://doi.org/10.1038/s41598-017-16006-z

J, G. H. C., R, B. J., R, C. I., Lawrence, H., David, L., F, M. J., Jules, P., Sherman, R., M, T. S., & Camilla, T. (2010). Food security: the challenge of feeding 9 billion people. Science, 327(5967), 812–818. https://doi.org/10.1126/science.1185383

Jones, J. D. G., & Dangl, J. L. (2006). The plant immune system. Nature, 444(7117), 323–329. https://doi.org/10.1038/nature05286

Jorge, D., & Jan, D. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. Science, 316(5833), 1862–1866. https://doi.org/10.1126/science.1143986

Lee, T. G., Kumar, I., Diers, B. W., & Hudson, M. E. (2015). Evolution and selection of *Rhg1*, a copy-number variant nematode-resistance locus. Molecular Ecology, 24(8), 1774–1791. https://doi.org/10.1111/mec.13138

Ling, H. Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., Li, Y., Yu, Y., Du, H., Qi, M., Li, Y., Lu, H., Yu, H., Cui, Y., Wang, N., Chen, C., … Liang, C. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature, 557(7705), 424–428. https://doi.org/10.1038/s41586-018-0108-0

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. PLoS Computational Biology, 13(5), e1005457–e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Luo, M. C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L., Wang, Y., McGuire, P. E., Liu, S., Long, H., Ramasamy, R. K., Rodriguez, J. C., Van Sonny, L., Yuan, L., Wang, Z., Xia, Z., … Dvoák, J. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature, 551(7681), 498–502. https://doi.org/10.1038/nature24486

Maccaferri, M., Harris, N. S., Twardziok, S. O., Pasam, R. K., Gundlach, H., Spannagl, M., Ormanbekova, D., Lux, T., Prade, V. M., Milner, S. G., Himmelbach, A., Mascher, M., Bagnaresi, P., Faccioli, P., Cozzi, P., Lauria, M., Lazzari, B., Stella, A., Manconi, A., … Cattivelli, L. (2019). Durum wheat genome highlights past domestication signatures and

future improvement targets. Nature Genetics, 51(5), 885–895. https://doi.org/10.1038/s41588-019-0381-3

Marone, D., Russo, M. A., Laidò, G., De Leonardis, A. M., & Mastrangelo, A. M. (2013). Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: Active guardians in host defense responses. International Journal of Molecular Sciences, 14(4), 7302–7326. https://doi.org/10.3390/ijms14047302

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., Gerhardt, D. J., Jeddeloh, J. A., & Stupar, R. M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiology, 159(4), 1295–1308. https://doi.org/10.1104/pp.112.194605

Miller and, W. A., & Rasochová, L. (1997). *Barley yellow dwarf viruses*. Annual Review of Phytopathology, 35(1), 167–190. https://doi.org/10.1146/annurev.phyto.35.1.167

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T., Mayer, K. F. X., Taudien, S., Platzer, M., Jeddeloh, J. A., Springer, N. M., Muehlbauer, G. J., & Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biology, 14(6), R58–R58. https://doi.org/10.1186/gb-2013-14-6-r58

Narang, D., Kaur, S., Steuernagel, B., Ghosh, S., Bansal, U., Li, J., Zhang, P., Bhardwaj, S., Uauy, C., Wulff, B. B. H., & Chhuneja, P. (2020). Discovery and characterisation of a new leaf rust resistance gene introgressed in wheat from wild wheat *Aegilops peregrina*. Scientific Reports, 10(1), 7573–7573. https://doi.org/10.1038/s41598-020-64166-2

Ogura, T., & Busch, W. (2015). From phenotypes to causal sequences: Using genome wide association studies to dissect the sequence basis for variation of plant development. Current Opinion in Plant Biology, 23, 98–108. https://doi.org/10.1016/j.pbi.2014.11.008

Ploetz, R. C. (2006). *Fusarium* wilt of banana is caused by several pathogens referred to as *Fusarium oxysporum* f. Sp. *cubense*. Phytopathology®, 96(6), 653–656. https://doi.org/10.1094/PHYTO-96-0653

Ploetz, R. C. (2015). *Fusarium* wilt of banana. Phytopathology®, 105(12), 1512–1521. https://doi.org/10.1094/PHYTO-04-15-0101-RVW

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome, 5(3). https://doi.org/10.3835/plantgenome2012.05.0005

Ramírez-González, R. H., Borrill, P., Lang, D., Harrington, S. A., Brinton, J., Venturini, L., Davey, M., Jacobs, J., Van Ex, F., Pasha, A., Khedikar, Y., Robinson, S. J., Cory, A. T., Florio, T., Concia, L., Juery, C., Schoonbeek, H., Steuernagel, B., Xiang, D., … Uauy, C. (2018). The transcriptional landscape of polyploid wheat. Science, 361(6403), 6089–eaar6089. https://doi.org/10.1126/science.aar6089

Richter, T. E., Pryor, T. J., Bennetzen, J. L., & Hulbert, S. H. (1995). New rust resistance specificities associated with recombination in the *Rp1* complex in maize. Genetics, 141(1), 373–381. https://doi.org/10.1093/genetics/141.1.373

Saintenac, C., Jiang, D., & Akhunov, E. D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biology, 12(9), R88–R88. https://doi.org/10.1186/gb-2011-12-9-r88

Sánchez-Martín, J., & Keller, B. (2019). Contribution of recent technological advances to future resistance breeding. Theoretical and Applied Genetics, 132(3), 713–732. https://doi.org/10.1007/s00122-019-03297-1

Shigenaga, A. M., Berens, M. L., Tsuda, K., & Argueso, C. T. (2017). Towards engineering of hormonal crosstalk in plant immunity. Current Opinion in Plant Biology, 38, 164–172. https://doi.org/10.1016/j.pbi.2017.04.021

Singh, K., & Kundu, J. K. (2018). *Wheat streak mosaic virus*. Plant Viruses, 131–148. https://doi.org/10.1201/b22221-8

Singla, J., & Krattinger, S. G. (2015). Biotic stress resistance genes in wheat. In Encyclopedia of Food Grains: Second Edition (Vols. 4–4, pp. 388–392). Elsevier Inc. https://doi.org/10.1016/B978-0-12-394437-5.00229-1

Thomas, M., R, S. S., Lise, H., Manuel, S., Matthias, P., S, J. K., H, W. B. B., Burkhard, S., X, M. K. F., Odd-Arne, O., Jane, R., Jaroslav, D., Curtis, P., Kellye, E., Catherine, F., Bikram, G., Bernd, F., J, L. A., … Sebastien, P. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. Science, 345(6194), 1250092–1250092. https://doi.org/10.1126/science.1250092

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D., … Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature, 588(7837), 277–283. https://doi.org/10.1038/s41586-020-2961-x

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484

Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Molecular Ecology, 28(6), 1203–1209. https://doi.org/10.1111/mec.15066

Winfield, M. O., Wilkinson, P. A., Allen, A. M., Barker, G. L. A., Coghill, J. A., Burridge, A., Hall, A., Brenchley, R. C., D'Amore, R., Hall, N., Bevan, M. W., Richmond, T., Gerhardt, D. J., Jeddeloh, J. A., & Edwards, K. J. (2012). Targeted re-sequencing of the allohexaploid wheat exome. Plant Biotechnology Journal, 10(6), 733–742. https://doi.org/10.1111/j.1467-7652.2012.00713.x

Xie, Y., Ravet, K., & Pearce, S. (2021). Extensive structural variation in the Bowman-Birk inhibitor family in common wheat (*Triticum aestivum* L.). BMC Genomics, 22(1), 218–218. https://doi.org/10.1186/s12864-021-07475-8

Yu, P., Wang, C.-H., Xu, Q., Feng, Y., Yuan, X.-P., Yu, H.-Y., Wang, Y.-P., Tang, S.-X., & Wei, X.-H. (2013). Genome-wide copy number variations in *Oryza sativa* L. BMC Genomics, 14(1), 649–649. https://doi.org/10.1186/1471-2164-14-649

Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2021). Current status of structural variation studies in plants. Plant Biotechnology Journal, n/a(n/a). https://doi.org/10.1111/pbi.13646

Zhu, T., Wang, L., Rimbert, H., Rodriguez, J. C., Deal, K. R., De Oliveira, R., Choulet, F., Keeble-Gagnère, G., Tibbits, J., Rogers, J., Eversole, K., Appels, R., Gu, Y. Q., Mascher, M., Dvorak, J., & Luo, M.-C. (2021). Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. The Plant Journal, 107(1), 303–314. https://doi.org/10.1111/tpj.15289

Zipfel, C. (2014). Plant pattern-recognition receptors. Trends in Immunology, 35(7), 345–351. https://doi.org/10.1016/j.it.2014.05.004

CHAPTER 2. EXTENSIVE STRUCTURAL VARIATION IN THE BOWMAN-BIRK INHIBITOR FAMILY IN COMMON WHEAT (*TRITICUM AESTIVUM* L.)[1]

2.1 Summary

Bowman-Birk inhibitors (BBI) are a family of serine-type protease inhibitors that modulate endogenous plant proteolytic activities during different phases of development. In this study, we used a Hidden Markov Model (HMM) profile-based search to identify 57 BBI genes in the common wheat (*Triticum aestivum* L.) genome. The BBI genes are unevenly distributed, with large gene clusters in the telomeric regions of homoeologous group 1 and 3 chromosomes that likely arose through a series of tandem gene duplication events. The genomes of wheat progenitors also contain contiguous clusters of BBI genes, suggesting this family underwent expansion before the domestication of common wheat. However, the BBI gene family varied in size among different cultivars, showing this family remains dynamic. Because of these expansions, the BBI gene family is larger in wheat than other monocots such as maize, rice and *Brachypodium*. We found BBI proteins in common wheat with intragenic homologous duplications of cysteine-rich functional domains, including one protein with four functional BBI domains. This diversification may expand the spectrum of target substrates. Expression profiling suggests that some wheat BBI proteins may be involved in regulating endogenous proteases during grain development, while others were induced in response to biotic and abiotic stresses. This information will facilitate the functional characterization of individual wheat BBI genes to determine their role in wheat development and stress responses and their potential application in breeding.

[1] Published in *BMC Genomics Journal*. Authors: Yucong Xie, Karl Ravet, and Stephen Pearce. (Xie et al., 2021)

2.2 Introduction

Plant proteases play vital roles in diverse biological processes by modulating programmed cell death, nutrient remobilization and defense responses (Clemente et al., 2019). Their activity is regulated by different classes of protease inhibitors (PIs) which bind to their protease substrates either through an irreversible trapping reaction or a tight-binding reaction (Laskowski & Kato, 1980; Laskowski & Qasim, 2000; S. Bateman & N.G. James, 2011). In plants, PIs regulate the activity of endogenous proteases to prevent proteolytic degradation, for example, by controlling the mobilization of storage proteins in seeds and kernels, and regulating senescence (Pak & Van Doorn, 2005; Volpicella et al., 2011). They also play important roles in plant defense by regulating the activity of exogenous proteases from different types of pests and pathogens to prevent cellular damage (Haq et al., 2004). In response to insect feeding, plant PIs are released into the insect's guts and inhibit digestive protease enzymes, which can prevent nutrient absorption, retarding their growth and development (Chen, 2008). Plant PIs are also induced by effector triggered immunity in response to bacterial and fungal pathogens to inhibit their proteolytic enzymes (Hellinger & Gruber, 2019; Jashni et al., 2015; Lawrence & Koundal, 2002). PIs are categorized into four broad classes according to their target protease specificity: serine PI (serpins), cysteine PI (cystatins), aspartic acid PI (pepstatins), and metallo-carboxy PI (Laskowski & Kato, 1980). PIs are further classified into types, families and clans to reflect their evolutionary relationships based on sequence homology, structural variation and biochemical function (Birk, 2003; Rawlings et al., 2004; Rawlings & Barrett, 1993). The latest PI classifications are maintained in the MEROPS database (Rawlings et al., 2018).

Bowman-Birk inhibitors (BBIs) are a family of serine-type PIs in MEROPS family I12, clan IF, that inhibit trypsin and chymotrypsin protease activity via the tight-binding reaction mechanism

(Birk et al., 1963; Bowman, 1946). Members of the BBI family are best known for their role in plant defense against phytophagous insects, and have been used to engineer insect-resistant transgenic crops (Singh et al., 2020). Overexpression of a cowpea trypsin inhibitor gene, which encodes a BBI protein, confers resistance to insects in the orders Coleoptera and Lepidoptera in tobacco (Hilder et al., 1987), rice (Xu et al., 1996), and wheat (Bi et al., 2006). Several BBI proteins also exhibit trypsin-like protease inhibition against fungal pathogens including *Mycosphaerella arachidicola, Fusarium oxysporum*, and *Botrytis cinerea* (Komarnytsky et al., 2006; Ye et al., 2001), *Fusarium culmorum* (Pekkarinen et al., 2007) and *Pyricularia oryzae* (Qu et al., 2003), as well as bacterial pathogens such as *Xanthomonas oryzae* pv. *Oryzae* (Pang et al., 2013). One rice BBI, APIP4, interacts at the protein level with both a fungal effector and host NLR receptors as part of the innate immune response, and plants carrying loss-of-function mutations in this gene exhibit increased susceptibility to *Magnaporthe oryzae* (Zhang et al., 2020). In wheat, genetic mapping studies identified putative BBI genes as candidates for seedling resistance to tan spot (Juliana et al., 2018) and *Fusarium* head blight (Sari et al., 2019). There is also evidence that BBIs play roles in more diverse processes, such as tolerance to salinity (Shan et al., 2008), oxidative (Dramé et al., 2013), and drought stress (Malefo et al., 2020; Yan et al., 2009), and regulating Fe uptake via an unknown mechanism (Zhang et al., 2014).

First discovered in soybean in 1946 (Bowman, 1944), BBIs had until recently only been described in the Fabaceae and Poaceae families (Mello et al., 2003). The BBIs are now known to be widely distributed in angiosperms (James et al., 2017; Mello et al., 2003; Qi et al., 2005), and evolutionary and phylogenetic analyses suggest they share a common ancestral sequence (James et al., 2017). The characterization of five BBIs in *Selaginella moellendorffii,* the oldest known extant vascular plant, show that this ancestral protein has a characteristic “double-headed”

structure with two homologous and spatially separated inhibitory loops within one BBI domain (James et al., 2017). Conserved inhibitory loops form reactive motifs providing dual specificity (Mello et al., 2003). BBI domains are also characterized by a series of conserved Cysteine (Cys) residues, which form disulfide bridges to provide structural stability required to maintain inhibitory loop conformation (Mello et al., 2003; Qi et al., 2005). The mutation of a single conserved Cys residue forming a disulfide bridge is sufficient to abolish the activity of either inhibitory loop (Clemente et al., 2015), and BBI domains with fewer than ten Cys residues are predicted to be non-functional (Mello et al., 2003). The Cys-formed inhibitory loops contain reactive domains composed of variable amino acids responsible for binding to trypsin and to chymotrypsin, including two residues, P1 and P1', that are proposed to play a role in determining protease substrate specificity (Mello et al., 2003). BBI proteins also commonly have a hydrophobic signal peptide (SP) at their N-terminus, with high sequence diversity among different BBIs (Baek et al., 1994; Baek & Kim, 1993). The SP is required for BBI protein translocation and secretion into the extracellular space, although it is not necessary for protease inhibition since the inhibitory loops can function independently of the rest of the BBI protein (Nishino et al., 1977). There is also evidence that BBI proteins can act in the nucleus (Zhang et al., 2014). All characterized BBI proteins in dicotyledonous plants have a conserved "double-headed" structure with a consistent molecular weight of approximately 8 kDa (Birk, 1985; James et al., 2017; Mello et al., 2003; Qi et al., 2005).

By contrast, almost all BBIs in monocotyledonous plants lack conserved Cys residues in the second inhibitory loop that are required to inhibit chymotrypsin, leading to a "single-headed" structure so that each BBI domain consists of only one functional reactive loop to inhibit trypsin activity (Mello et al., 2003). The only known exceptions are three "double-headed" BBIs in the

banana (*Musa acuminate*) genome, indicating that the "single-headed" BBI structure originated since the monocot and dicot lineages diverged (James et al., 2017). Evolutionary models indicate that monocot BBIs underwent internal domain duplications within a single protein that resulted in multiple inhibitory loops (Mello et al., 2003; Prakash et al., 1997; Qu et al., 2003). Previous studies divided monocot BBI proteins into six groups (MI-I to MI-VI) on the basis of their functional domain number and the number and position of conserved Cys residues (Habib & Fazili, 2007; Lawrence & Koundal, 2002; Mello et al., 2003). To simplify, these six BBI models in monocots can be grouped into three broad classes; one comprised of 8 kDa proteins with a single functional domain (groups MI-I, MI-II, and MI-III), a second class with a molecular weight of approximately 16 kDa and a duplicated single-inhibitory loop (groups MI-IV and MI-V) and a final category of larger proteins with three tandemly duplicated BBI domains. While the first two classes are widespread in monocots, only three rice BBIs have been described which fall into the final class (Qu et al., 2003).

Genome-wide studies of the BBI gene family have been performed in rice (Qu et al., 2003), common bean (Galasso et al., 2009) and other angiosperms (James et al., 2017). However, to date, only three BBIs have been characterized in common wheat (*Triticum aestivum* L.), a crop which provides approximately 20% of the calories and proteins consumed by the human population (*USDA-FAO*, 2018). Of the three BBI proteins isolated from wheat germ, IBB1 has two homologous functional domains, each with one functional inhibitory loop (Odani et al., 1986; Raj et al., 2002), whereas IBB2 and IBB3 have only one functional domain (Odani et al., 1986; Poerio et al., 1994). These three BBIs inhibit protease activity, control protein metabolism during wheat kernel development and germination, and inhibit fungal trypsin-like activity and hyphal growth (Chilosi et al., 2000). Three other putative genes with sequence homology to BBIs (*wali3, wali5,*

and *wali6*) were isolated as cDNAs from wheat root tips (Richards et al., 1994; Shan et al., 2008; Snowden et al., 1995). These putative BBI genes are transcriptionally induced by wounding or by the imposition of toxic metal stress, but their function against protease was not tested (Shan et al., 2008; Snowden et al., 1995).

The identification of wheat BBI genes is complicated by the high frequency of residue substitution and sequence variability among encoded proteins, and the complexity of the wheat genome. Common wheat is an allopolyploid (genomes AABBDD) produced from two separate hybridization events. The first occurred approximately 0.5 to 0.9 million years ago between *T. urartu* (AA) and an unknown species related to *Aegilops speltoides* to form the tetraploid wild emmer wheat *T. turgidum* ssp. *dicoccoides* (AABB). A second hybridization event between *T. turgidum* ssp. durum and *Ae. tauschii* (DD) gave rise to common wheat, approximately 10,000 years ago (El Baidouri et al., 2017).

In the current study, we used a Hidden Markov Model (HMM)-based approach to describe the BBI gene family in common wheat, revealing it to be larger than in other monocot species. We found evidence of extensive gene duplications throughout wheat's evolutionary history, as well as internal duplications that further diversified the functional BBI domains of individual proteins. The findings from our study highlight the extent of variation in the BBI gene family in the Triticeae lineage and will facilitate their functional characterization to explore how this diversity impacts wheat development and plant defense.

2.3 Results

2.3.1 Bowman-Birk inhibitor genes are unevenly distributed in the common wheat genome

We identified 57 BBI genes in the hexaploid common wheat genome using a three-step HMM-based approach outlined in Figure 2.1. We first used the HMM profile for BBI (Pfam:

PF00228, downloaded from the Pfam database) to search the IWGSC RefSeq v1.1 protein database and identified 39 BBI proteins. We generated a new HMM profile based on the alignment of these 39 sequences and used this in a second search against the same protein database to identify 62 BBI proteins, including 23 that were not found in the first step. We performed HMMscan on each protein and excluded five sequences that lacked a BBI Pfam domain (Table S2.1). A final search using an HMM profile built from an alignment of the remaining 57 BBIs did not yield any additional proteins, confirming this is a comprehensive list of annotated BBI proteins in the wheat landrace 'Chinese Spring' (Table S2.1).

We manually adjusted the start codon position for five BBIs to match homologous sequences (Table S2.2). After manual curation, 50 full-length BBIs are predicted to have an N-terminal SP domain, with cleavage positions ranging from 15 to 30 amino acids. Seven N-terminally truncated BBIs are predicted to lack a functional SP domain (Table S2.1).

The 57 BBIs include three genes (*TraesCS3A02G046000*, *TraesCS3B02G036400*, and *TraesCS1B02G025900*) that encode previously characterized BBI proteins - IBB1, IBB2, and IBB3 (Table S2.3) (Odani et al., 1986; Poerio et al., 1994). Three other previously described putative BBI genes (*wali3, wali5* and *wali6* (Richards et al., 1994; Snowden et al., 1995)) were not found among the 57 BBIs. An HMMscan analysis of the corresponding full-length proteins (*TraesCS1D02G265900*, *TraesCS1D02G265800* and *TraesCS1B02G276900*) revealed that they did not contain a BBI domain, indicating these genes do not encode functional BBI proteins (Table S2.3).

Wheat BBI genes are unevenly distributed across the genome with two gene triads on chromosomes 4 and 5 and large clusters on homoeologous group 3 (36 BBIs) and group 1 chromosomes (15 BBIs) (Figure 2.2). The BBI genes in these clusters are separated by short

physical distances and in several instances include adjacent BBIs, suggesting they arose through tandem gene duplication events (Figure 2.2). For example, the ten BBIs on chromosome 3A span a region of just 270 kb and include four adjacent BBIs (Figure 2.2B). All wheat BBIs were located in the telomeric regions (R1 and R3) of their respective chromosomes (Figure 2.2A).

This pattern of gene duplication is consistent with homology analysis that divided the 57 BBIs into six homoeologous categories (Table 2.1). Overall, 21 BBI genes (36.8% of the total) formed seven complete triads (1:1:1 for A:B:D genome), close to the 35.8% for all wheat genes in the genome (Appels et al., 2018). By contrast, 14% of BBI genes form groups characterized by gene duplication (n:1:1/1:n:1/1:1:n) compared to 5.7% of all wheat genes (Appels et al., 2018) (Table 2.1). In addition, one group of genes consisted of four tandemly duplicated genes on chromosome 1B (0:4:0), while on chromosome 3, one group exhibited duplications of both the A and B homoeologs (2:2:1) (Table 2.1, Table S2.4).

Table 2.1 Homoeologous group identification and categorization of the BBI gene family in wheat

Category number	Homoeologous group (A:B:D)	Number of groups	Number of genes	% of genes
1	1:1:1	7	21	36.8
2	2:1:1 and 1:2:1	2	8	14
3	1:1:0 and 0:1:1	2	4	7
4	0:4:0	1	4	7
5	2:2:1 and 2:0:2	2	9	15.8
6	Singletons	11	11	19.4
-	Total	25	57	100

To determine whether these duplication events affected the selective pressure on BBI genes, we performed a Ka/Ks ratio analysis to calculate the sequence divergence rate for the clusters of BBIs on individual homoeologous group 1 and 3 chromosomes. A ratio of non-synonymous (Ka) to synonymous (Ks) nucleotide changes greater than one indicates divergent function of two genes,

whereas a Ka/Ks ratio of less than one indicates purifying selection and conserved function. The Ka/Ks ratios for pairwise comparisons of BBI genes on homoeologous group 1 chromosomes were all less than one, except for one branch on chromosome 1D between *TraesCS1D02G020600* and *TraesCS1D02G018700LC* that had a value of 1.17 (Figure S2.1). By contrast, eight branches on homoeologous group 3 chromosomes had Ka/Ks values greater than one, including four branches on 3A, two branches on 3B, and two branches on 3D (Figure S2.1).

Overall, our analysis shows that the BBI family in wheat is unevenly distributed across the genome and includes large gene clusters in the telomeric regions of homoeologous group 1 and group 3 chromosomes. The distribution of the genes in these clusters suggest they originated from paralogous expansion through tandem duplication events.

2.3.2 BBI genes underwent extensive tandem duplications in the Triticeae

We next compared the BBI family in wheat with other monocot species. Using the same approach and criteria (Figure 2.1), we identified six BBIs from *Brachypodium* (*B. distachyon*), seven from maize (*Z. mays*), eleven from rice (*O. sativa*), and sixteen from barley (*H. vulgare*) (Figure 2.3A). A full list of BBIs from each species is provided in Table S2.5. Considering its hexaploid genome, common wheat has an average of 19 BBI genes per diploid genome, 3.2-fold more than *Brachypodium*, 2.7-fold more than maize, 1.7-fold more than rice, but just 1.2-fold more than barley (Figure 2.3B).

To explore the genetic relationships between BBIs in these species, we constructed a phylogenetic tree from all identified proteins. The tree separated wheat BBIs into three broad clades, each of which also contained BBIs from other species, except clade A that does not contain maize BBIs (Figure 2.3C). Clade A clustered all wheat BBIs located on homoeologous group 1 and 5 chromosomes. Clade B included the majority of wheat BBIs located on homoeologous group

3 chromosomes, with the remainder clustered in clade C together with the BBI gene triad from chromosome 4 (Figure 2.3C).

Consistent with their relatively recent divergence and the similarity in size of the BBI gene family, most barley BBIs co-located with wheat BBIs (Figure 2.3C). However, one cluster of contiguous BBIs on barley chromosome 3H suggests that gene duplication events also occurred independently in this species (Clade C, Figure 2.3C). Maize and rice BBIs formed two distinct clusters in clade B and clade C, which included several adjacent BBIs in their respective genome assemblies, suggesting that BBI gene duplication also occurred independently in both these species (Figure 2.3C).

BBI proteins were also separated according to the type of reactive site and the number of active domains they contained, as defined by Mello *et al.* (Mello et al., 2003). Every BBI from all species in clade A contains a single active BBI domain and all fall into the MI-I group except for one barley BBI (*HORVU5Hr1G068510*) that does not match any previously characterized BBI group (Figure 2.3C). The wheat BBIs clustered in clades B and C are all multi-domain proteins and fall into either the MI-II or MI-IV groups except for three wheat BBIs with more than two domains that are most similar to the MI-IV group (Figure 2.3C). The cluster of rice, maize and *Brachypodium* BBIs in clade C were most similar to the wheat BBIs on homoeologous group 3 chromosomes, and were also all multi-domain proteins, represented by groups MI-IV, MI-V and MI-VI (Figure 2.3C).

This phylogeny reveals that the BBI gene family in monocots is subject to a complex pattern of internal and external gene duplication events, resulting in multi-domain BBIs and gene copy number variation in each species. In wheat, extensive gene duplication on homoeologous group 1

and especially group 3 chromosomes, that also occurred in barley, account for the greater numbers of BBI genes in the Triticeae lineage compared to other grasses.

2.3.3 The BBI gene family underwent gene duplication and deletion events both before and after common wheat's domestication

To gauge the approximate timing of the BBI gene family expansion in wheat, we identified BBI proteins from common wheat's ancestors. We found 12 BBIs from *T. urartu*, and 17 from *Ae. tauschii*, the diploid progenitors of the A and D genomes of common wheat, respectively (Figure 2.4A). Because the diploid wheat B genome progenitor is unknown, we analyzed *T. dicoccoides*, an allotetraploid progenitor with genomes AABB, and identified 23 BBIs. We excluded one of these genes from our analysis (*TRIDCUv2G007850*) because it was not assembled into a known chromosome, leaving eight BBIs on the A genome and fourteen on the B genome (Figure 2.4A). Compared to each diploid progenitor genome, the corresponding genome in *T. aestivum* contained a greater number of BBIs (Figure 2.4B). There were 1.3-fold more BBIs on the A genome of *T. aestivum* than in *T. urartu* and 1.9-fold more genes than in the A genome of *T. dicoccoides* (Figure 2.4B). There were 1.5-fold more BBIs on the B genome of *T. aestivum* compared to *T. dicoccoides*. By contrast, the *T. aestivum* D genome contains only 1.2-fold more BBI genes than *Ae. tauschii* (Figure 2.4B).

Phylogeny showed that most genes from wheat ancestors were clustered into orthologous groups with their corresponding genes in common wheat (Figure 2.4C, Table S2.6). Orthologs of the BBI genes on *T. aestivum* homoeologous group 4 chromosomes were present in *T. urartu* (A genome) and *Ae. tauschii* (D genome), but absent from *T. dicoccoides* (AB genomes). Orthologs of the BBI genes on *T. aestivum* group 5 chromosomes were present in both A and B genomes of

T. dicoccoides and in the D genome of *Ae. tauschii* (Table S2.6). None of these genes were duplicated in any wheat species.

By contrast, the similarity and genomic position of BBI gene clusters on homoeologous group 1 and 3 chromosomes in progenitor wheat species suggests that many BBI gene duplication events occurred before common wheat's domestication. On *Ae. tauschii* chromosome 1D, six contiguous BBI genes are clustered within 800 kb, while on chromosome 3D, eight BBI genes are clustered within 500 kb, suggesting they arose through tandem duplication (Table S2.5). In *T. dicoccoides*, there are five BBI genes on chromosome 3A within a 264 kb region and thirteen BBI genes on chromosome 3B within a 696 kb region (Table S2.5). This phylogeny also revealed several instances of gene duplications in hexaploid *T. aestivum* that were absent in the diploid or tetraploid progenitors. For example, we found a cluster of four adjacent paralogous BBIs on chromosome 1B of *T. aestivum* that were all absent from *T. dicoccoides,* suggesting that tandem duplication events occurred after common wheat's domestication (Figure 2.4C and Table S2.6).

To analyze the diversity within the BBI gene family arising from selections made during domestication and breeding, we identified BBIs in the genome assemblies of four common wheat cultivars (Table S2.7). The total number of BBI genes in these cultivars ranged from 55 in 'Mace' to 60 in 'Jagger' (Table 2.2). While the BBI gene triads on chromosomes 4 and 5 were conserved in all cultivars, phylogenetic analysis indicated several instances of gene loss and gain on homoeologous group 1 and 3 chromosomes (Figure S2.2). Although the BBI gene number varied between cultivars on each of these chromosomes, this variation was greatest on chromosomes 1B, 1D and 3B (Figure 2.5, Table 2.2). Strikingly, none of the five analyzed cultivars shared an identical complement of BBI genes.

Table 2.2 BBI genes in five common wheat varieties, separated by chromosome.

Chromosome	Chinese Spring	Jagger	Mace	Julius	Landmark	$\Delta_{max-min}$
1A	3	3	3	3	2	1
1B	5	5	4	6	3	3
1D	7	9	9	9	10	3
3A	10	11	10	10	11	1
3B	14	13	11	12	13	3
3D	12	13	12	13	13	1
4A	1	1	1	1	1	0
4B	1	1	1	1	1	0
4D	1	1	1	1	1	0
5A	1	1	1	1	1	0
5B	1	1	1	1	1	0
5D	1	1	1	1	1	0
Total	57	60	55	59	58	5

$\Delta_{max-min}$ shows the inter-varietal variation in BBI gene number for each chromosome.

Taken together, analysis of the BBI gene family in different wheat germplasm reveals that while the gene triads on chromosomes 4 and 5 did not undergo expansion throughout wheat evolution, the gene clusters on homoeologous group 1 and 3 chromosomes are more variable. Many gene duplication events occurred before domestication, but the increase in gene number in common wheat and variation among modern wheat cultivars shows that the BBI family remains dynamic.

2.3.4 Wheat BBI genes on homoeologous group 3 chromosomes encode proteins with duplicated active domains

We next studied in greater detail the functional domains in the 57 BBIs from 'Chinese Spring'. The majority of wheat BBIs (36 proteins, 63%) had one functional BBI domain, including all 15 BBIs located on homoeologous group 1 chromosomes, the gene triads on chromosomes 4 and 5 and 15 BBIs on homoeologous group 3 chromosomes (Figure 2.6AB). Of the remaining BBIs on group 3 chromosomes, 18 had two functional BBI domains (Figure 2.6C), two proteins

(TraesCS3D02G036400 and TraesCS3D02G035700) had three domains and one protein (TraesCS3B02G038300) had four domains (Figure 2.6D). The gene structure of wheat BBIs reveals that while the majority have either one (6 BBIs, 10%) or two exons (45 BBIs, 79%), five genes including all three-domain proteins had three exons, while the gene (*TraesCS3B02G038300*) encoding the four-domain protein had four exons (Figure S2.3). This suggests that the genes encoding three- or four-domain BBI proteins may have evolved either from complete or partial gene duplication followed by fusion of tandem-duplicated genes. However, the genomic sequences encoding these domains, including intron and flanking sequences, were variable between domains, suggesting they did not arise from recent duplication events.

We characterized the number and positions of conserved Cys residues within the reactive motifs of wheat BBIs according to the evolutionary scheme of Mello *et al*. (Mello et al., 2003) and with respect to substrate specificity. The first reactive inhibitory motif for trypsin was predicted to be conserved and functional in all wheat BBIs except for three proteins: TraesCS3A02G049400LC, which has a truncated motif in a functional BBI domain (Figure 2.6B), TraesCS3D02G033700, which has a two amino-acid deletion within the reactive motif (Figure 2.6B), and TraesCS3B02G037200, which carries a Cys to Tyrosine (Y) amino acid substitution in the final residue of the first reactive motif (Figure 2.6C). The vast majority of wheat BBIs carried K/R-S amino acids at the P1 and P1' positions, respectively (Figure 2.6), the consensus motif for monocot trypsin inhibition (Mello et al., 2003). All MI-I type BBIs had a K/R-S motif except one protein (TraesCS1D02G019800LC) that has a Glycine (G) in the P1 residue (Figure 2.6A). Among the MI-II type BBIs, the P1-P1' motif was more diverse. Notably, four homoeologous BBIs on group 3 chromosomes each exhibited Serine (S) to Valine (V) substitutions at position P1', while each protein in the triad on chromosome 4 had a Glutamate (E) residue at position P1 (Figure 2.6B).

Most MI-IV type BBIs also had a K/R-S motif in both reactive sites except a homoeologous triad of BBIs with Serine to Tyrosine (T) substitutions in the P1' residue (Figure 2.6C). In all 57 wheat BBIs the disulfide bridge (C_{10} and C_{11}) supporting the second inhibitory motif for chymotrypsin was lost (Figure 2.6).

The gene triad on chromosome 5 and all 15 BBIs on homoeologous group 1 chromosomes each encode BBI proteins with functional domains comprised of 12 Cys residues that form six disulfide bridges, except for TraesCS1A02G022000 and TraesCS1A02G019800LC which carry amino acid substitutions at Cys residues in positions C_6 and C_{14}, respectively (Figure 2.6A). The homoeologous triad of BBIs on chromosome 4 fall into the MI-II group (Figure 2.6B). BBIs on group 3 chromosomes were the most divergent. There were 15 BBIs categorized into the MI-II group that each contain ten Cys residues, except for TraesCS3A02G049400LC which has a deletion encompassing four Cys residues, and TraesCS3A02G045700, TraesCS3B02G042700LC and TraesCS3D02G034100 which each carry a single Cys amino acid substitution (Figure 2.6B). Another 18 BBIs encode two-domain proteins categorized in the MI-IV group although three (TraesCS3D02G035300, TraesCS3B02G037300 and TraesCS3B02G03740) had a truncated second domain, while another protein (TraesCS3B02G03720) has fewer than ten Cys residues in both domains (Figure 2.6C). The three wheat BBIs with more than two domains could not be categorized into any previously described MI evolutionary group (Figure 2.6D). The three-domain proteins TraesCS3B02G036400 and TraesCS3D02G035700 are most similar to the MI-IV group but each underwent internal duplication of one domain resulting in three adjacent BBI domains that have distinct Cys positions from the previously proposed MI-VI three-domain group (Mello et al., 2003). TraesCS3B02G036400 has a truncated second domain and a deletion of five Cys residues while the Cys positions in TraesCS3D02G035700 are also divergent from existing models

(Figure 2.6D). Each of the four domains in TraesCS3B02G038300 are full length and contain ten conserved Cys residues, suggesting all four may be functional (Figure 2.6D). In summary, five BBIs (TraesCS3A02G045700, TraesCS3D02G034100, TraesCS3B02G042700LC, TraesCS3A02G049400LC, and TraesCS3B02G03720) have fewer than ten Cys residues in one or both BBI domains, and are predicted to be non-functional. While four other multi-domain BBIs (TraesCS3B02G036400, TraesCS3B02G03730, TraesCS3B02G03740, and TraesCS3D02G035300) have fewer than ten Cys residues in one domain, these proteins are predicted to exhibit protease inhibition activity since at least one other domain remains intact (Figure 2.6D and Table S2.1). A summary of the different types of wheat BBI proteins in common wheat is shown in Figure 2.7.

2.3.5 Wheat BBI genes exhibit diverse expression profiles during development and in response to biotic and abiotic stress

We next used public RNA-seq datasets to characterize transcript levels of the 57 BBI genes in common wheat (Borrill et al., 2016). Genes were clustered into four main groups based on their expression profile in different wheat tissues and at different stages of development (Figure 2.8A). Genes in group I showed relatively high transcript levels in most plant tissues during development. BBIs in group II were predominantly expressed in root tissues, while BBIs in group III were expressed most highly during the early stages of leaf, stem and spike development. Finally, genes in group IV showed low levels of expression in most tissues and included ten genes with no detectable transcripts in any assayed tissue (Figure 2.8A).

We also identified a subset of wheat BBIs that exhibit stress-responsive changes in expression (Figure 2.8B). The majority of the highly expressed BBIs in group I are induced in response to stripe rust and *Septoria tritici* blotch infection and are suppressed by heat stress (Figure 2.8B).

Several BBIs in other groups were induced by multiple biotic stresses, including some genes in group IV that were only expressed in response to stress, indicating they may play a role in general immunity. Many of the BBI genes with no detectable expression in any of the reported conditions encode proteins lacking a SP or with truncations and amino acid changes in critical domains, suggesting they may be non-functional (Figure 2.8).

2.4 Discussion and Conclusion

2.4.1 Diverse wheat genomic resources facilitate gene family characterization studies

In this study, we identified and characterized the BBI gene family in the common wheat landrace 'Chinese Spring', four modern cultivars, and their extant progenitors, using HMM-based homology searches (Figure 2.1). This approach incorporates position-specific alignment scores and ensemble algorithms to evaluate all possible alignments. By weighting the relative likelihood of each alignment to identify orthologous proteins against a Pfam protein database, HMM may provide greater sensitivity than other sequence-based searches to identify all members of a gene family (Potter et al., 2018). In addition, Pfam annotations are more specific than superfamily protein groupings that are assembled in other databases, allowing for the more stringent classification of proteins. For each species, a single HMMsearch using a profile downloaded from the Pfam database was insufficient to identify all BBI proteins, likely because this general profile does not reflect species-specific diversity in this protein family (El-Gebali et al., 2019). A second search using a custom HMM profile built from an alignment of BBIs from the first screen yielded additional BBIs in every species analyzed, and for wheat, included 13 BBI proteins not associated with a BBI Pfam domain in their IWGSC RefSeq v1.0 gene model annotations (Appels et al., 2018). We confirmed that each protein contained at least one BBI Pfam domain using HMMscan, although it is important to note that these sequences represent *in silico* predictions and the

inhibitory function of each protein should be validated using biochemical assays, especially for those lacking conserved Cys residues.

Access to a greater diversity of high-quality genome assemblies for wheat will allow for more detailed gene characterization studies in this species. For example, the recent assembly of a more contiguous 'Chinese Spring' genome using both short and long read sequencing resolved 5,799 gene duplications that were not annotated in IWGSC RefSeq v1.1 (Alonge et al., 2020). These include two BBI genes (*T4033720*, a paralog of *TraesCS3A02G046300* that is located 6 Mb downstream on the same chromosome, and *T4042195*, a paralog of *TraesCS5B02G498100* located on chromosome 3B) that were not present in the IWGSC RefSeq v1.1 assembly (Table S2.8). Beyond 'Chinese Spring', an international wheat pan-genome project aims to sequence and assemble multiple common wheat genomes (Walkowiak et al., 2020). Among the five varieties we analyzed in this study, no two had the same complement of BBI genes (Figure 2.5), although it is important to note that presence/absence variation between varieties may be the result of incomplete genome assembly. A set of fully-annotated, high-quality genome assemblies of diverse wheat varieties will be a valuable resource to characterize the full extent of natural genetic variation in wheat.

2.4.2 The wheat BBI gene family underwent extensive duplication resulting in copy number variation and multi-domain proteins

Consistent with previous studies, our phylogenetic analysis shows that the BBI family is subject to widespread gene duplication events that likely occurred independently in each monocot species since they last shared a common ancestor (James et al., 2017; Qu et al., 2003). In rice, ten BBI genes are located in a 430 kb region of chromosome 1 (Qu et al., 2003), in maize, four BBIs are 200 kb apart on chromosome 3 and in barley, eleven BBIs are located within a 450 kb region

of chromosome 3H (Figure 2.3C, Table S2.5). Each of these regions is syntenic with the distal region of wheat homoeologous group 3 chromosomes (La Rota & Sorrells, 2004; Munkvold et al., 2004), suggesting that a common mechanism associated with this region of the genome, likely conserved in all crop species, triggers gene duplication at these loci. We found some evidence of other gene duplication events within 200 kb of BBI gene clusters in wheat, including 11 genes encoding proteins annotated as "Disease resistance RPM1" on chromosome 1B and 12 genes encoding E3-Ubiquitin ligase proteins on chromosome 3D (Table S2.9 and Table S2.10). However, these duplications were not shared between homoeologous chromosomes, suggesting they arose from recent duplication events, and are unlikely to be associated with BBI protein function in wheat.

One possible factor contributing to the high rate of duplications in the BBI family may be the location of gene clusters in distal telomeric regions of each chromosome (Figure 2.2), which are hotspots for evolution, recombination events (Glover et al., 2015) and, in polyploid species, homoeologous exchange (Zhang et al., 2020). Characterization of the MADS-box transcription factor family in wheat revealed a positive correlation between the number of genes in a subfamily and their proximity to the telomere (Schilling et al., 2020). In barley, large segmental duplications occurred more frequently in the telomeres, and were associated with increased gene copy number variation, potentially because of higher rates of non-allelic homologous recombination in these regions (Bretani et al., 2020). However, their position alone cannot account for the extent of BBI duplication, because the genes on homoeologous group 4 and 5 chromosomes are similarly located in the telomere but did not undergo duplication in any barley or wheat genome analyzed in our study (Figure 2.3C).

Although we found evidence of BBI gene duplication in all analyzed monocot genomes, this family was larger in wheat and barley due to more extensive tandem duplication events on wheat homoeologous group 1 and 3 chromosomes and barley chromosome 3H (Figure 2.3). Although many of these gene duplication events had already occurred in wheat's diploid and tetraploid progenitors, we also identified several duplication events that occurred since common wheat's domestication (Figure 2.4), demonstrating that the process driving BBI family expansion in wheat remains active. In polyploid wheat species, relaxed selection pressure arising from gene redundancy may partially account for the greater expansion of the BBI gene family (Comai, 2005). However, the similar size of the BBI gene family in diploid barley and wheat progenitors shows that gene duplication occurs to a similar degree in different Triticeae species, demonstrating that polyploidy is not necessary for BBI duplication. Further studies will be required to determine the mechanism or factors driving BBI gene family expansion in the Triticeae.

Our study also revealed that BBI domain duplication, possibly originating from incomplete gene duplication followed by gene fusion or internal duplication, resulted in further diversification of encoded wheat BBI proteins, potentially enlarging the spectrum of their protease substrates (Figure 2.6). Despite the high level of conservation of the P1-P1' motif in wheat (Figure 2.6), this motif is more variable in other monocots such as rice and banana (James et al., 2017; Qu et al., 2003), so its importance for substrate recognition will require further analysis. Domain duplication is a common feature of BBI evolution in different plant species, including an ancient event that gave rise to the "double-headed" BBI structure conserved in dicots (James et al., 2017; Mello et al., 2003; Qu et al., 2003). Our *in silico* analysis predicted that all wheat BBI proteins lack a functional second reactive motif to inhibit chymotrypsin activity (Figure 2.6), consistent with analyses of other monocot BBIs (Mello et al., 2003; Qi et al., 2005). However, previous studies

have detected chymotrypsin inhibition in protein extracts from the wheat endosperm, so it is likely that this activity is performed by a distinct family of protease inhibitors, potentially members of the cereal trypsin/α-amylase inhibitor family (Di Maro et al., 2011; Tedeschi et al., 2012).

Multi-domain monocot BBIs were previously isolated and characterized in other monocot species (Mello et al., 2003; Nagasue et al., 1988; Qu et al., 2003; Tashiro et al., 1990). The separation of all single-domain BBIs and all multi-domain BBIs in our phylogenetic tree suggests that these multi-domain BBIs were already present in the common ancestor of these grasses (Figure 2.3C). In wheat, all multi-domain BBIs are located on homoeologous group 3 chromosomes (Figure 2.6). Our finding that BBIs on both group 1 and group 3 chromosomes underwent complete gene duplication but only the BBIs on group 3 chromosomes underwent domain duplication (Figure 2.3C and Figure 2.6), suggests that the mechanism of gene duplication differs between group 1 and group 3 chromosomes. Alternatively, the reduced selective pressure on BBI genes on homoeologous group 3 chromosomes (Figure S2.1) may result in a higher magnitude of gene expansion and an increased frequency of internal duplications giving rise to multi-domain proteins. These include three- and four-domain BBI proteins distinct in structure from any previously proposed BBI protein model (Figure 2.6D). In order to determine the impact of this variation, it will be critical to identify the endogenous and exogenous interacting substrates of the BBI family, which remain poorly understood.

2.4.3 Functional characterization of wheat BBI genes

Gene duplication events can impact molecular evolution in different ways (Magadum et al., 2013). These include: (i) loss of protein function resulting from excessive mutation accumulation (ii) gain of protein function as a result of gene overexpression, (iii) neo- or sub-functionalization, and (iv) modulation of protein activity by duplicating and diversifying reactive sites. Our analyses

indicate that the wheat BBI family potentially contains members exhibiting each of these features. Several wheat BBI genes exhibited truncations, mutations in active sites and undetectable transcript levels in all assayed tissues (Figure 2.8A), suggesting they may be non-functional pseudogenes. Conversely, we also identified homoeologous BBI genes that exhibit divergent expression profiles, suggesting they may have taken on new functional roles during wheat development (Figure 2.8A). Several wheat BBIs exhibit high transcript levels in the grain, suggesting they may regulate endogenous protease activity during grain development (Figure 2.8A). We also identified a subset of BBIs that are transcriptionally induced in response to fungal and bacterial pathogens, consistent with previous studies in other plants (Chilosi et al., 2000; Othman et al., 2014; Qu et al., 2003), which may indicate these genes contribute to plant defense responses (Figure 2.8B). One of these genes, *TraesCS1A02G021400*, is induced in response to four pathogens (Figure 2.8B) and was previously identified as a candidate gene for wheat seedling resistance to tan spot (Juliana et al., 2018). It would be interesting to characterize this gene to determine its potential role in disease resistance in wheat. Another wheat BBI gene, *TraesCS1B02G025900*, was identified as a candidate defense hub gene for Type II Fusarium head blight resistance (Sari et al., 2019). However, this BBI gene is expressed primarily in root and stem tissues and is not induced in response to any biotic stress assayed in our study (Figure 2.8), suggesting it is unlikely to play a role in disease resistance. Some pathogens secrete proteases as part of their infection cycle, and in response, plants have co-evolved different classes of PIs to inhibit their activity (Qu et al., 2003). In wheat, a greater number of BBI proteins with more numerous and diverse reactive sites may allow the wheat plant to inhibit a wider range of pathogenic protease substrate variants as part of an effective response against fungal and bacterial pathogens (Qu et al., 2003). Identification of the protease inhibitors interacting with wheat BBIs

will allow for a more detailed understanding of their mode of action. Future studies might include an analysis of the co-expression of BBIs and their protease targets during development and in response to biotic or abiotic stress. It is interesting to note that the distal area of chromosome arm 3BS, which includes a cluster of 14 BBI genes, overlaps with the Wheat Streak Mosaic Virus resistance locus *Wsm2* (Dhakal et al., 2018; Tan et al., 2017). Although there is no evidence that BBI proteins act as an R gene for virus resistance, they might function as antagonistic interacting proteins with other R proteins to trigger defense responses (Malefo et al., 2020).

In conclusion, we found that the BBI gene family in common wheat is larger than in other monocots due to a series of tandem duplication events in the telomeric regions of homoeologous group 1 and group 3 chromosomes. The increased frequency of gene duplications on homoeologous group 3 chromosomes likely gave rise to multi-domain BBI proteins with novel reactive sites. It will be important to determine the endogenous and exogenous protease substrates of individual BBIs and to identify how divergent and duplicated active sites impact their specificity and activity. Our description of this gene family in wheat will facilitate the functional characterization of individual BBI genes. Reverse genetics tools will facilitate hypothesis testing to determine the role of BBI genes in wheat development and defense responses (Uauy et al., 2017), and to help identify natural genetic variation that may be valuable for elite cultivar development.

2.5 Methods

2.5.1 Identification of Bowman-Birk inhibitors in plant genomes

High and low confidence wheat protein annotations from IWGSC RefSeq v1.1 (Appels et al., 2018) were downloaded from the IWGSC sequence repository hosted by URGI (https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.1/) and concatenated into a single FASTA file consisting of 298,774 protein sequences. Protein sequences

were obtained from the reference assemblies of *Hordeum vulgare* (IBSC_v2, 236,301 protein sequences), *Brachypodium distachyon* (v3.0, 52,972 protein sequences), *Aegilops tauschii* (Aet_v4.0, 258,680 protein sequences) (Luo et al., 2017), and *Triticum urartu* (ASM34745v1, 33,483 protein sequences) (Ling et al., 2018) from Ensembl Plants (https://plants.ensembl.org/info/website/ftp/index.html). *Oryza sativa* proteins were downloaded from the Rice Genome Annotation Project (Oryza_japonica.MSUv7, 55,986 protein sequences) (RGAP, http://rice.plantbiology.msu.edu) and converted to IRGSP-1.0 gene IDs and *Zea mays* proteins (Zea_mays.B73_RefGen_v4, 131,585 protein sequences) were downloaded from MaizeGDB (https://www.maizegdb.org). *Triticum turgidum* ssp. *dicoccoides* wild emmer wheat 'Zavitan' WEWseq v2 proteins (205,916 sequences) were downloaded from https://search.datacite.org/works/10.5447/ipk/2019/0 (Avni et al., 2017).

The identification of BBI proteins in each species was performed with HMMER analysis (Potter et al., 2018) against the local protein annotation database using a three-step approach outlined in Figure 2.1. First, we performed an HMMsearch using the HMM profile for the Bowman-Birk protease inhibitor family (Pfam: PF00228) which was downloaded from Pfam 32.0 (El-Gebali et al., 2019) using an E-value threshold of $1e^{-5}$. We next aligned the BBI protein sequences identified from the first step using HMMalign and built a new HMM profile based on the multiple alignment using HMMbuild. We used the new generated HMM profile to conduct a second HMMsearch against the same species-specific protein databases. Finally, we examined the list of BBI proteins for the presence of a BBI Pfam domain (PF00228) using HMMscan with an E-value threshold of 0.05. Proteins that contained the Pfam domain were classified as BBI. We then performed alignment of the identified BBI protein sequences from all species with MAFFT (Katoh et al., 2018) and noticed that several BBIs were predicted to lack a signal peptide due to

misannotation of the methionine start codon. We manually curated the position of the N-terminal start codon of several BBIs from *T. aestivum*, *Ae. tauchii*, *T. urartu*, *T. dicoccoides* and *H. vulgare* to match homologous sequences. Full curation details are provided in Table S2.2, and includes details of BBI proteins with N-terminal truncations likely caused by point mutations. The curated sequences were used in all subsequent analyses.

2.5.2 Chromosomal locations and homology identification

All identified wheat BBIs were mapped to the IWGSC RefSeq v1.1 genome assembly to identify their chromosomal location (Appels et al., 2018). To determine homologous relationships between genes, we performed all-to-all BLAST using the 57 proteins as queries and applied an E-value threshold of $1e^{-10}$. Putative paralogs or homoeologs were defined as homologous BBIs with a BLASTP e-value < $1e^{-10}$ and identity > 75% on the same or homoeologous group chromosome, respectively. This approach was also used to identify orthologous relationships between BBIs in common wheat and progenitor genomes. The synteny and homologous relationship of wheat BBI genes were visualized with Circos plot using R shinyCircos (Y. Yu et al., 2018). Each chromosome was divided into telomere (R1/R3), centromere (C) and R2 segments according to information from the IWGSC RefSeq v1.1 genome assembly (Appels et al., 2018). The distance of wheat BBIs and other high- and low-confidence gene models were mapped to individual chromosomes using the R Sushi package plotBed function (Phanstiel et al., 2014).

We calculated Ka/Ks ratios using an online tool hosted by the computational biology unit (CBU http://services.cbu.uib.no/tools/kaks) using the coding sequence of each common wheat BBI gene. We excluded one BBI (TraesCS3D02G035800) from the Ka/Ks ratio analysis due to a premature termination codon in its coding sequence. The remaining 56 BBIs were grouped

according to their chromosome and used to construct phylogenetic trees and calculate the pairwise Ka/Ks ratio for each branch.

2.5.3 Alignment and phylogenetic analysis

We performed multiple sequence alignments using Clustal Omega using full-length BBI protein sequences identified in all species. Model selection was conducted with IQ-TREE using the lowest Bayesian information criterion (BIC) as WAG+G4 model (Kalyaanamoorthy et al., 2017). We constructed the phylogenetic tree using the selected model with 1000 ultrafast bootstrap replicates UFBoot2 (Hoang et al., 2018; Nguyen et al., 2015). The resulting tree was visualized and annotated with the R package ggtree v2.0.4 (G. Yu et al., 2017). The domain model type for BBIs in grasses were determined manually by comparing the number and position of Cys residues to the model proposed by Mello *et al.* (Mello et al., 2003).

2.5.4 Identification of BBI on homoeologous group 3 chromosomes in different wheat varieties

The draft genome assembly for four common wheat varieties 'Jagger' (U.S.A, winter growth habit), 'Julius' (Germany, winter), 'Landmark' (Canada, spring), and 'Mace' (Australia, spring) were downloaded from the 10+ Wheat Genomes Project (https://wheat.ipk-gatersleben.de/downloads/) and used to build local BLAST databases. We then used the full-length protein-coding sequences of each BBI gene from 'Chinese Spring' as queries and performed BLAST against the genomes of each wheat variety to identify their chromosomal position (Deng et al., 2007). The position of each BBI genes in these varieties was cross-referenced with GFF files to identify the corresponding gene ID provided by the 10+ Wheat Genome Project (Walkowiak et al., 2020). To identify BBIs present in these varieties but absent from the 'Chinese Spring' assembly, the corresponding genomic region spanning all BBI genes on homoeologous group 1

and group 3 chromosomes from each wheat variety were extracted locally using the bedtools getfasta command for *ab initio* gene prediction (Quinlan & Hall, 2010). The open reading frame (ORF) and putative gene model for each extracted DNA fragment was predicted with OrfM-0.7.1 (Woodcroft et al., 2016). To determine whether predicted gene models contain a functional BBI domain, all predicted ORFs were scanned with HMMscan using an e-value cutoff as 0.05 and those BBIs with PF00228 domains were retained. After exclusion of common orthologs in other varieties, the unique BBIs in each variety were named using the first two letters of the cultivar name, followed by chromosome number, and ordered by their relative position on that chromosome. For example, JA_3A-1 represents the first unique BBI on 'Jagger' chromosome 3A. The BBI protein sequences from all varieties were then used to construct a phylogenetic tree with IQ-TREE using the WAG+G4 model with 1000 ultrafast bootstrap replicates UFBoot2 (Hoang et al., 2018; Nguyen et al., 2015). The resulting tree was visualized and annotated with the R package ggtree v2.0.4 (G. Yu et al., 2017).

2.5.5 Gene structure analysis of the functional domains and motifs

The complete genomic, CDS and amino acid sequences, as well as gene feature information of all BBIs identified were downloaded from IWGSC RefSeq v1.1 (Appels et al., 2018). Schematic representation of the exon-intron organization of wheat BBIs was conducted by comparing the CDS and the corresponding genomic sequences using Gene Structure Display Server 2.0 (Hu et al., 2015). To find conserved Cys-rich domains, the amino acid sequence for the functional domains of all identified BBIs in wheat, by aligning amino acid sequences between the first and last conserved Cys residue in each domain using MAFFT v7 for multiple sequence alignment (Katoh et al., 2018). All sequences were analyzed using Signal P v5.0 (Almagro Armenteros et al., 2019) to predict the presence of N-terminal SP and for potential cleavage sites.

2.5.6 Gene expression analysis

The expression data for wheat BBI genes in five tissues (spike, root, leaf, grain and stem) at three different developmental stages from hexaploid wheat var. 'Chinese Spring' (Lukaszewski et al., 2014) and under abiotic stress (heat and drought) condition at the one-week-old seedling stage (Liu et al., 2015) were mapped to the IWGSC RefSeq v1.1 genome and processed into TPM values as described previously (Pearce et al., 2015). Separately, we downloaded several biotic stress expression datasets as TPM from the online wheat expression browser expVIP (Borrill et al., 2016), including studies on fusarium head blight (Kugler et al., 2013; Schweiger et al., 2016), stripe rust (Cantu et al., 2013; H. Zhang et al., 2014), powdery mildew (H. Zhang et al., 2014), fusarium crown rot (Powell et al., 2017), *Septoria tritici* blotch (Rudd et al., 2015; Yang et al., 2013) and PAMP elicitors (Ramírez-González et al., 2018). For each pathogen, we calculated the $\log_2$ fold change of the transcript abundance for each treated sample compared to mock controls or samples at time zero at each time point and averaged the values of all time points. Heatmaps for tissue specific time course expression were constructed using log2 transformed TPM values with the R package pheatmap v1.0.12. Genes were clustered according to their expression level (metric, Euclidian; method, complete) and grouped by their chromosome type.

CHAPTER 2 FIGURES

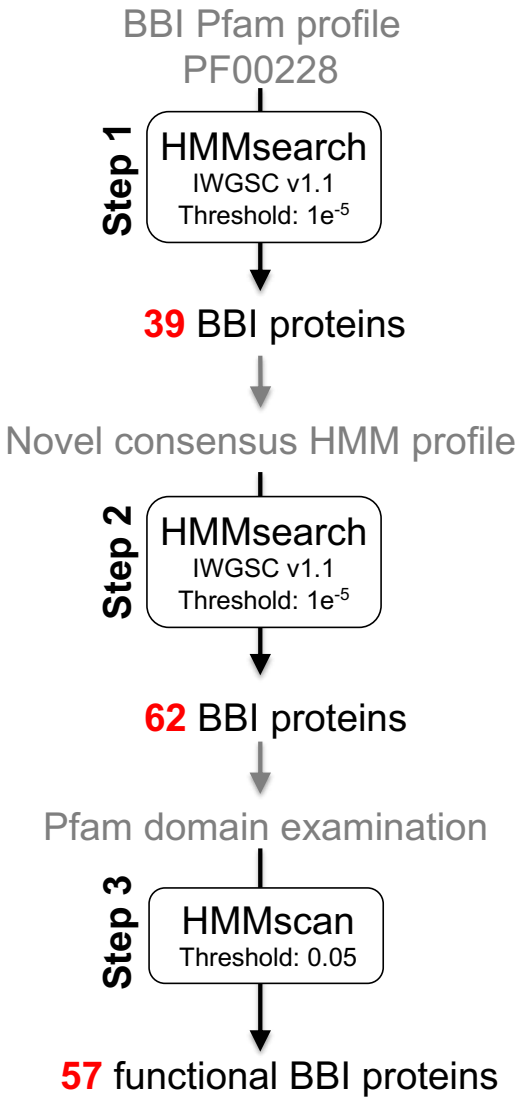


Figure 2.1 Pipeline for Bowman-Birk inhibitor (BBI) gene family identification in plant genomes. The identification of BBIs in the *T. aestivum* genome is presented as an example, including key steps and criteria for each step. The number of proteins identified at each stage are highlighted in red.

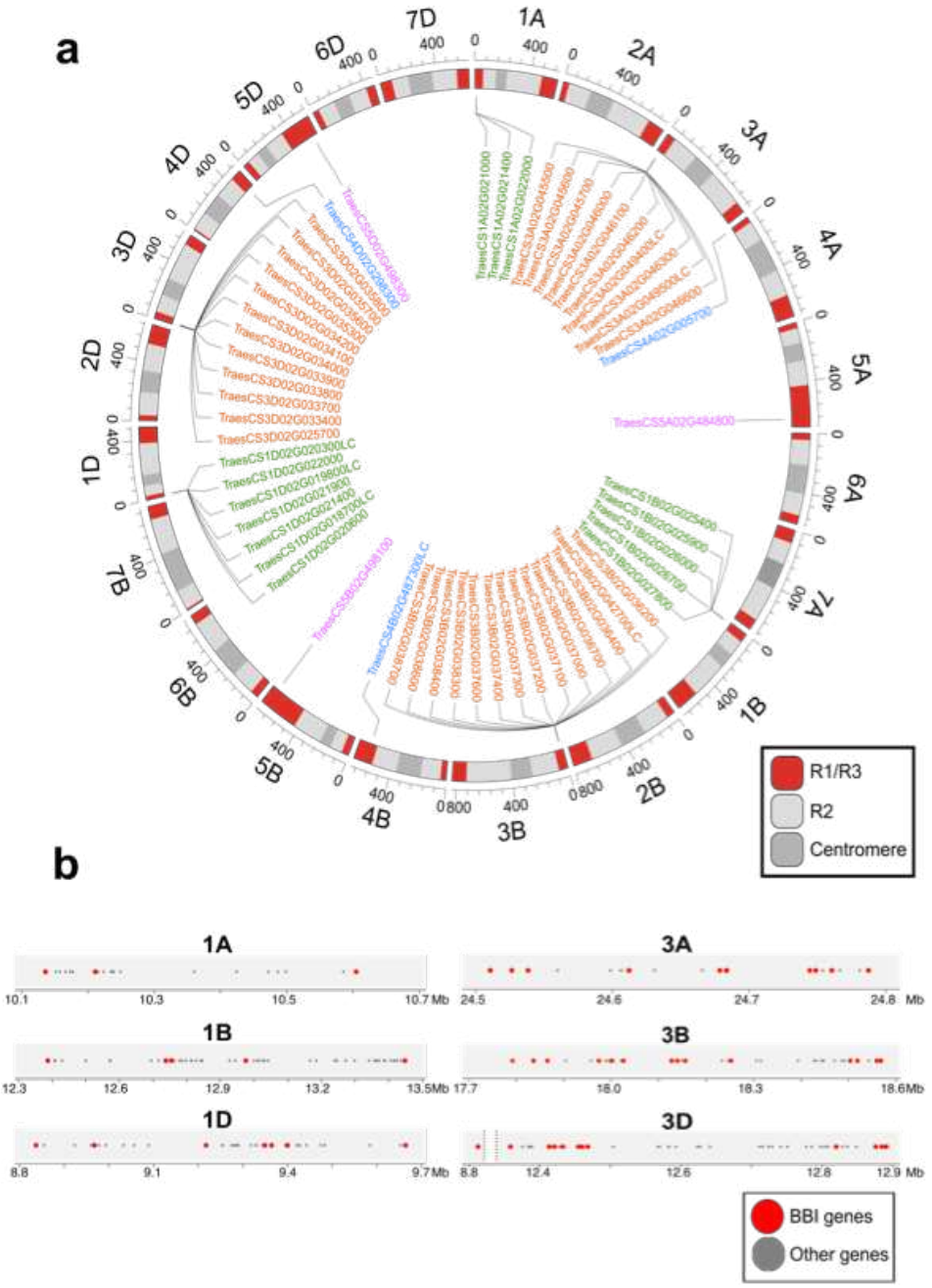


Figure 2.2 Distribution of 57 BBI in the *T. aestivum* genome **a** Chromosomal positions of wheat BBIs. Gene names are colored according to their homoeologous group. Chromosomal segments are indicated by different colors - distal regions of the chromosome R1 and R3 in red, centromeric region C in dark grey, and region R2 in light grey. **b** Distribution of genes within BBI clusters on homoeologous group 1 and group 3 chromosomes. Red dots represent BBI genes, whereas grey dot represent other annotated genes in the region, positioned according to their physical location in the IWGSC RefSeq v1.1 genome assembly. All high confidence (HC) and low confidence (LC) gene models are presented.

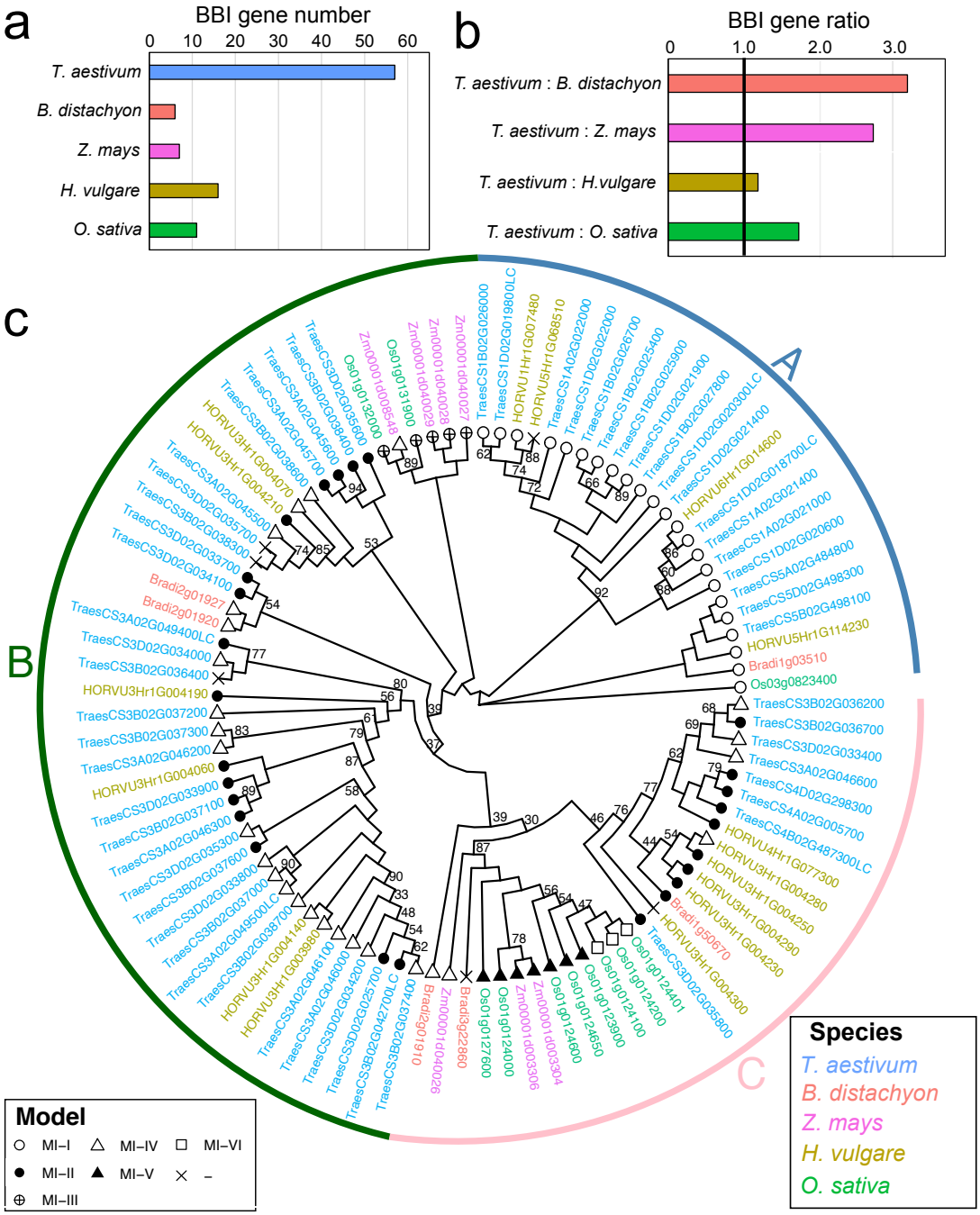


Figure 2.3 Comparison of the wheat BBI gene family with other monocots. **a** Total number of BBI genes in monocot genomes. Bars are color-coded based on species. **b** Ratios of total BBI gene numbers in common wheat compared to other monocot species, adjusted for wheat's hexaploid genome. The 1:1 ratio is indicated by a bold line. **c** Circular phylogenetic tree of all BBI proteins from rice, maize, barley, *Brachypodium* and common wheat. Only bootstrap support values below 95 are indicated on the tree. Gene labels are color-coded by species and includes the BBI group based on the classification of Mello *et al*. (Mello et al., 2003).

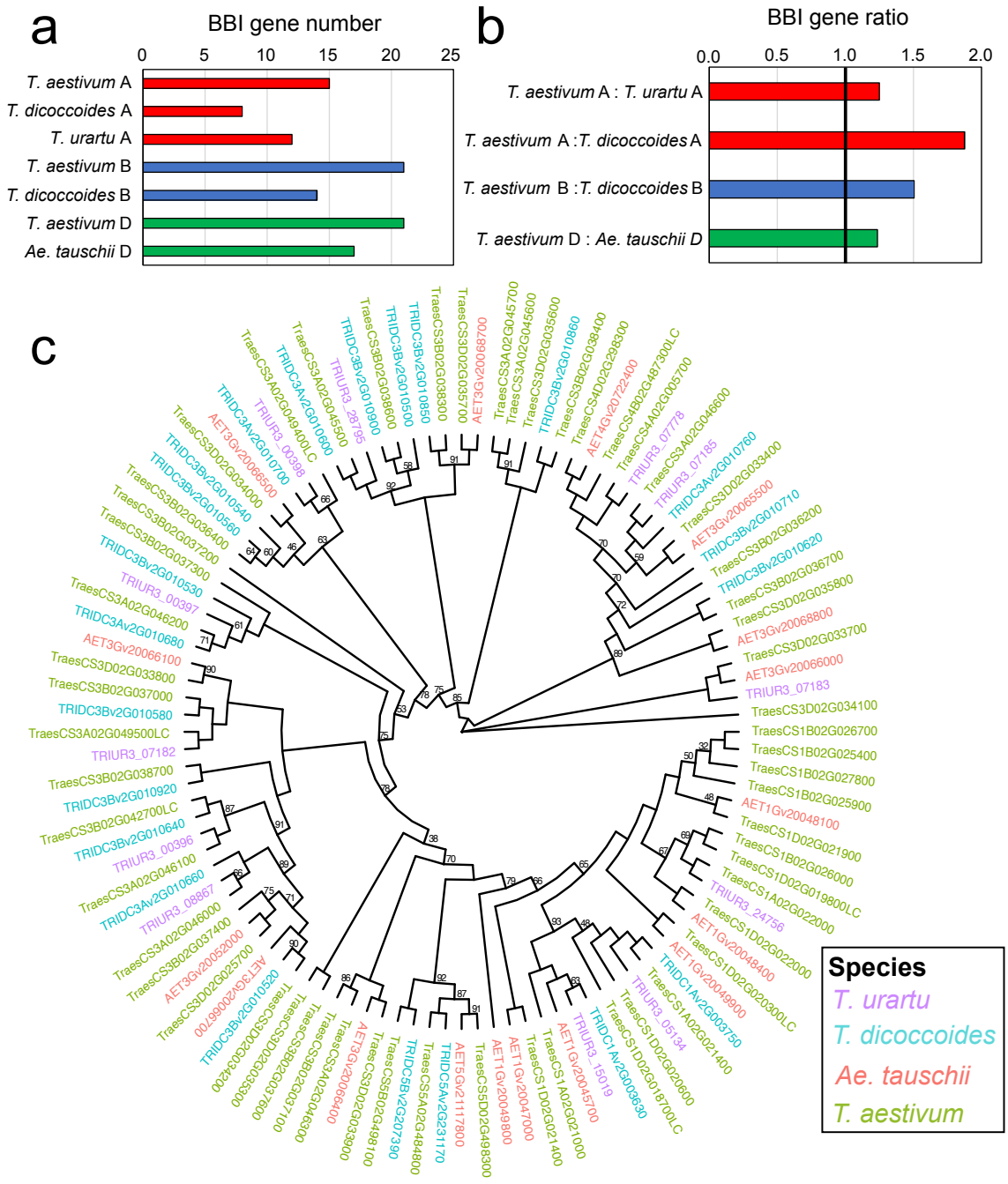


Figure 2.4 Comparison of the BBI gene family in different wheat germplasm. **a** The number of BBI genes in the genomes of different wheat species. Bars are color coded by species. **b** Ratio of total BBI gene numbers in common wheat compared to progenitor species. The 1:1 ratio is indicated by a bold line. **c** Phylogenetic tree constructed from all BBI proteins from each wheat species. Only bootstrap support values below 95 are indicated on the tree. Genes are color-coded based on species.

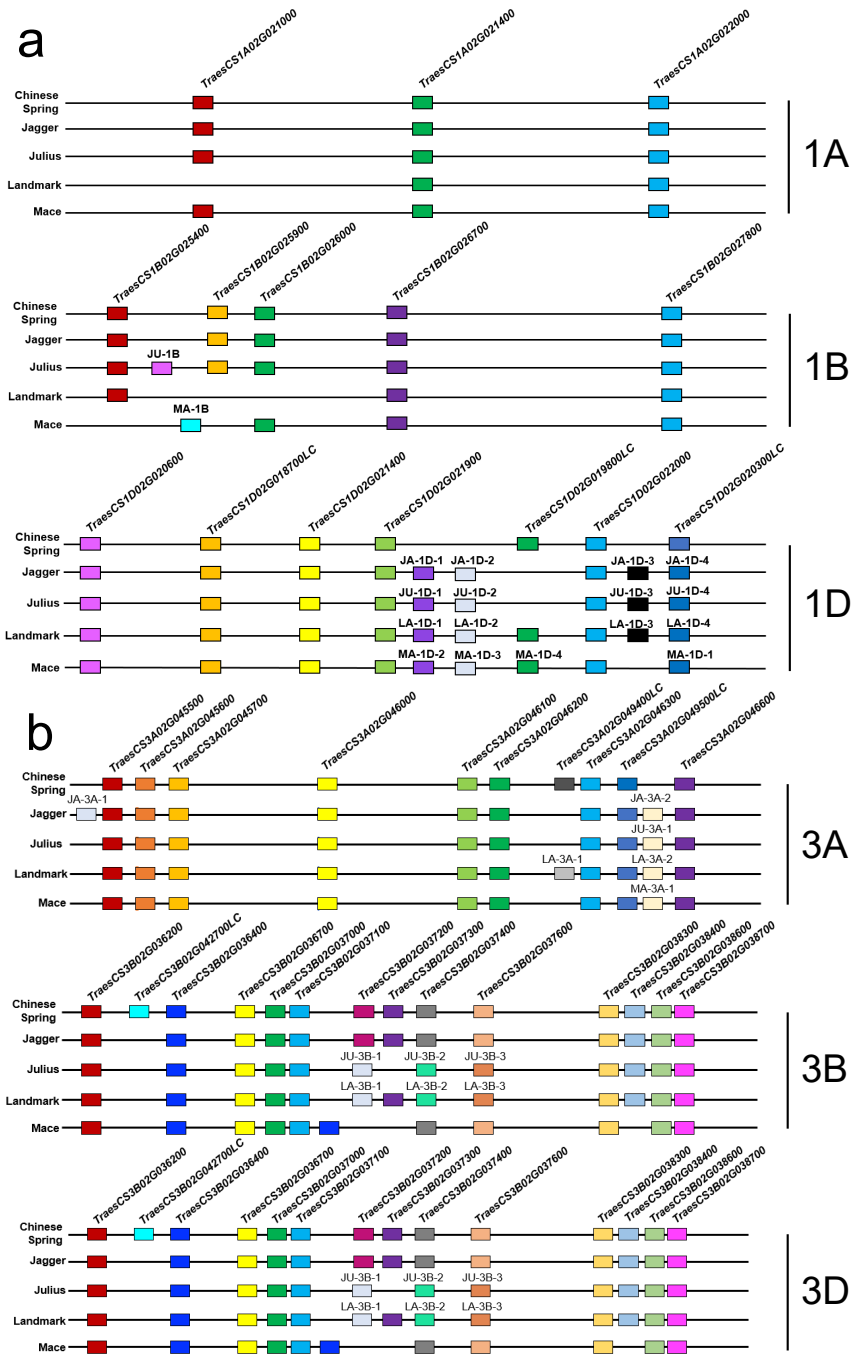


Figure 2.5 Distribution of BBI genes on homoeologous group 1 and 3 chromosomes in different common wheat varieties. **a**. Homoeologous group 1 chromosomes. **b**. Homoeologous group 3 chromosomes. The 'Chinese Spring' BBIs are ordered according to their physical position in the IWGSC RefSeq v.1.1 genome assembly, but not to scale. Genes are colored according to their homology so that genes in the same color are orthologous in different varieties. The BBIs present in other varieties but absent in 'Chinese Spring' are labeled such that JA_3A-1 indicates the first unique BBI on 'Jagger' chromosome 3A.

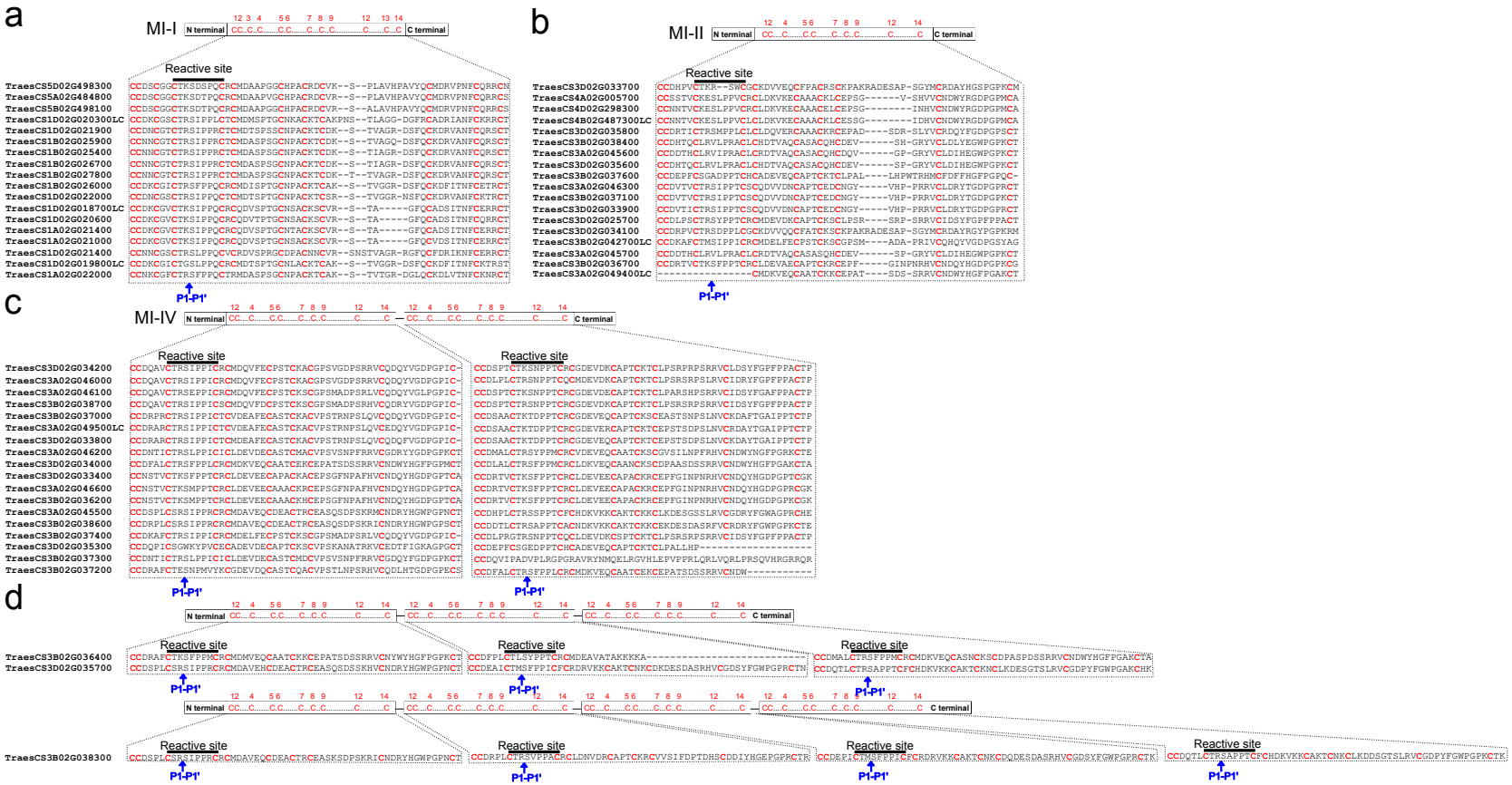


Figure 2.6 Alignment of the conserved Cys-rich domains of common wheat BBI proteins. **a** Alignment of common wheat BBI proteins falling into group MI-I; **b** MI-II group; **c** MI-IV group; and **d** BBI proteins that cannot be classified into an existing group. The Cys residues are highlighted in red with their corresponding position indicated above each alignment. The blue arrow underneath the domain sequences highlight the P1 and P1' positions.

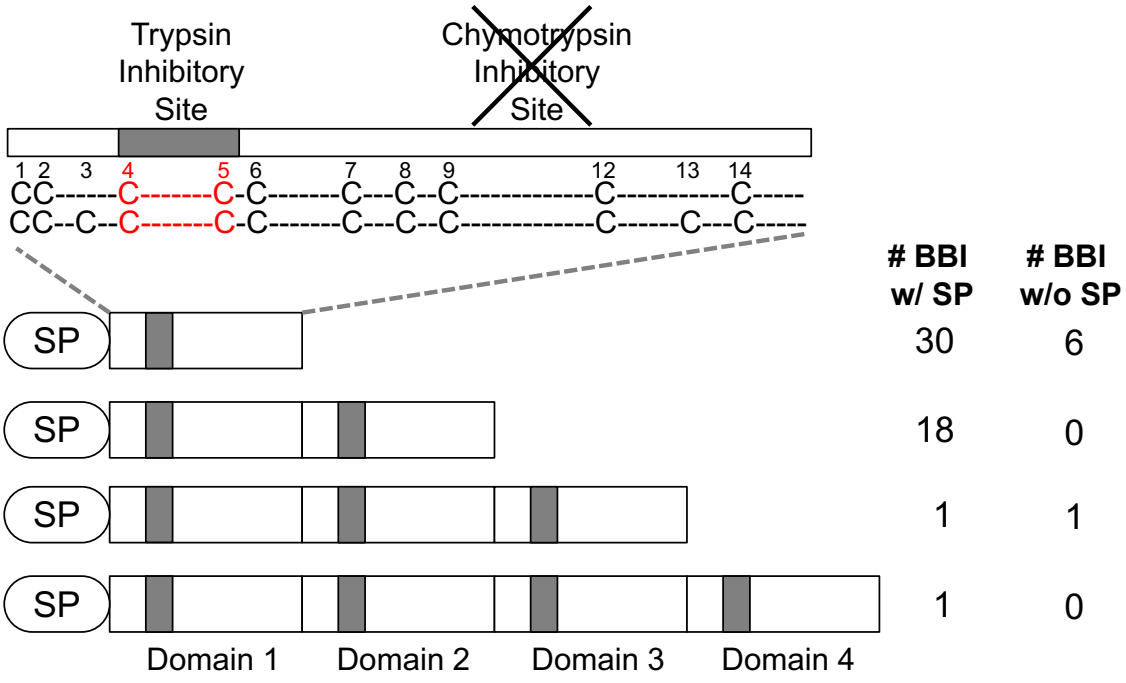


Figure 2.7 Summary of the proposed structural composition of the BBI gene family in common wheat. Most wheat BBI proteins contain an N-terminal signal peptide, and between one and four reactive loop domains at the C-terminus. The conserved Cys residues in the inhibitory domain are listed as C, and other amino acid residues indicated as dashes. The first reactive site is highlighted in red. The numbers of wheat BBIs from 'Chinese Spring' falling into each category are separated on the basis of presence or absence of complete signal peptides.

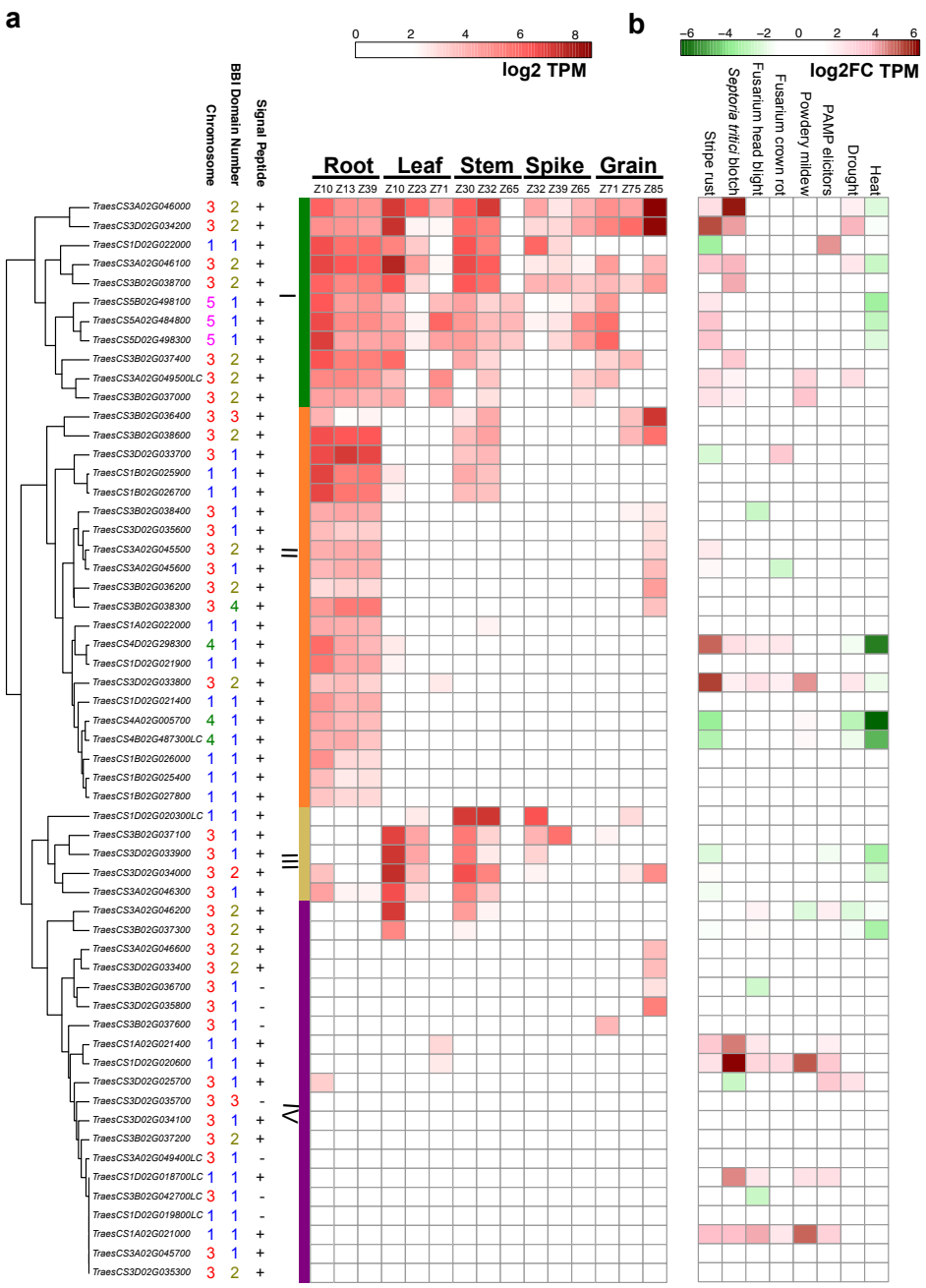


Figure 2.8 Expression profiles of common wheat BBI genes. **a** Transcript levels of BBI genes in five tissue types (root, leaf, stem, spike and grain) each at three different developmental stages based on the Zadoks scale. Genes were clustered based on their expression profile and all expression data is presented as log2 TPM. **b** Transcript levels of BBI genes in response to biotic and abiotic stress. Expression levels of each BBI gene in wheat plants infected with fusarium head blight, stripe rust, powdery mildew, fusarium crown rot, *Septoria tritici* blotch, and bacterium chitin and flg22 (PAMP) were compared to the corresponding mock treatment. Expression is presented as log2 fold change of the TPM between pathogen treatment and mock control. Drought and heat stress data is taken from one-week-old seedlings and presented as log2 fold change of the TPM between stress and control seedlings. Where multiple timepoints were in each dataset, mean fold-changes are presented. Genes are in the same orders as in the tissue specific developmental conditions in panel a. The number of BBI domains and chromosome for each gene are color-coded. The presence (+) or absence (-) of a signal peptide for each protein is indicated.

REFERENCES

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

Alonge, M., Shumate, A., Puiu, D., Zimin, A. V., & Salzberg, S. L. (2020). Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies. Genetics, 216(2), 599–608. https://doi.org/10.1534/genetics.120.303501

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A. A., Himmelbach, A., Balfourier, F., … Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 361(6403), 7191. https://doi.org/10.1126/science.aar7191

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., Jordan, K. W., Golan, G., Deek, J., Ben-Zvi, B., Ben-Zvi, G., Himmelbach, A., Maclachlan, R. P., Sharpe, A. G., Fritz, A., … Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science, 357(6346), 93–97. https://doi.org/10.1126/science.aan0032

Baek, J. M., & Kim, S. I. (1993). Nucleotide sequence of a cDNA encoding soybean Bowman-Birk proteinase inhibitor. Plant Physiology, 102(2), 687. https://doi.org/10.1104/pp.102.2.687

Baek, J. M., Song, J. C., Choi, Y. Do, & Kim, S. Il. (1994). Nucleotide sequence homology of cDNAs encoding soybean Bowman-Birk type proteinase inhibitor and its isoinhibitors.

Bioscience, Biotechnology, and Biochemistry, 58(5), 843–846. https://doi.org/10.1271/bbb.58.843

Bi, R. M., Jia, H. Y., Feng, D. S., & Wang, H. G. (2006). Production and analysis of transgenic wheat (*Triticum aestivum* L.) with improved insect resistance by the introduction of cowpea trypsin inhibitor gene. Euphytica, 151(3), 351–360. https://doi.org/10.1007/s10681-006-9157-9

Birk, Y. (1985). The Bowman-Birk inhibitor. Trypsin- and chymotrypsin-inhibitor from soybeans. In International Journal of Peptide and Protein Research (Vol. 25, Issue 2, pp. 113–131). Int. J. Peptide protein Res. https://doi.org/10.1111/j.1399-3011.1985.tb02155.x

Birk, Y. (2003). Plant protease inhibitors (1st ed.). Springer-Verlag Berlin Heidelberg.

Birk, Y., Gertler, A., & Khalef, S. (1963). A pure trypsin inhibitor from soybeans. The Biochemical Journal, 87(2), 281–284. https://doi.org/10.1042/bj0870281

Borrill, P., Ramirez-Gonzalez, R., & Uauy, C. (2016). ExpVIP: A customizable RNA-seq data analysis and visualization platform. Plant Physiology, 170(4), 2172–2186. https://doi.org/10.1104/pp.15.01667

Bowman, D. E. (1944). Fractions derived from soybeans and navy beans which retard tryptic digestion of casein. Proceedings of the Society for Experimental Biology and Medicine, 57(1), 139–140. https://doi.org/10.3181/00379727-57-14731P

Bowman, D. E. (1946). Differentiation of soybean antitryptic factors. Proceedings of the Society for Experimental Biology and Medicine, 63(3), 547–550. https://doi.org/10.3181/00379727-63-15668

Bretani, G., Rossini, L., Ferrandi, C., Russell, J., Waugh, R., Kilian, B., Bagnaresi, P., Cattivelli, L., & Fricano, A. (2020). Segmental duplications are hot spots of copy number variants

affecting barley gene content. Plant Journal, 103(3), 1073–1088. https://doi.org/10.1111/tpj.14784

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., Dubcovsky, J., Saunders, D. G. O., & Uauy, C. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. Sp. *Tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. BMC Genom, 14(1), 270. https://doi.org/10.1186/1471-2164-14-270

Chen, M. (2008). Inducible direct plant defense against insect herbivores: A review. Insect Science, 15(2), 101–114. https://doi.org/10.1111/j.1744-7917.2008.00190.x

Chilosi, G., Caruso, C., Caporale, C., Leonardi, L., Bertini, L., Buzi, A., Nobile, M., Magro, P., & Buonocore, V. (2000). Antifungal activity of a Bowman-Birk-type trypsin inhibitor from wheat kernel. Journal of Phytopathology, 148(7–8), 477–481. https://doi.org/10.1046/j.1439-0434.2000.00527.x

Clemente, A., Arques, M. C., Dalmais, M., Le Signor, C., Chinoy, C., Olias, R., Rayner, T., Isaac, P. G., Lawson, D. M., Bendahmane, A., & Domoney, C. (2015). Eliminating anti-nutritional plant food proteins: The case of seed protease inhibitors in pea. PLoS ONE, 10(8), e0134634–e0134634. https://doi.org/10.1371/journal.pone.0134634

Clemente, M., Corigliano, M. G., Pariani, S. A., Sánchez-López, E. F., Sander, V. A., & Ramos-Duarte, V. A. (2019). Plant serine protease inhibitors: Biotechnology application in agriculture and molecular farming. International Journal of Molecular Sciences, 20(6), 1345. https://doi.org/10.3390/ijms20061345

Comai, L. (2005). The advantages and disadvantages of being polyploid. Nature Reviews Genetics, 6(11), 836–846. https://doi.org/10.1038/nrg1711

Deng, W., Nickle, D. C., Learn, G. H., Maust, B., & Mullins, J. I. (2007). ViroBLAST: A stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. Bioinformatics, 23(17), 2334–2336. https://doi.org/10.1093/bioinformatics/btm331

Dhakal, S., Tan, C. T., Anderson, V., Yu, H., Fuentealba, M. P., Rudd, J. C., Haley, S. D., Xue, Q., Ibrahim, A. M. H., Garza, L., Devkota, R. N., & Liu, S. (2018). Mapping and KASP marker development for wheat curl mite resistance in "TAM 112" wheat using linkage and association analysis. Molecular Breeding, 38(10), 119. https://doi.org/10.1007/s11032-018-0879-x

Di Maro, A., Farisei, F., Panichi, D., Severino, V., Bruni, N., Ficca, A. G., Ferranti, P., Capuzzi, V., Tedeschi, F., & Poerio, E. (2011). WCI, a novel wheat chymotrypsin inhibitor: Purification, primary structure, inhibitory properties and heterologous expression. Planta, 234(4), 723–735. https://doi.org/10.1007/s00425-011-1437-5

Dramé, K. N., Passaquet, C., Repellin, A., & Zuily-Fodil, Y. (2013). Cloning, characterization and differential expression of a Bowman-Birk inhibitor during progressive water deficit and subsequent recovery in peanut (*Arachis hypogaea*) leaves. Journal of Plant Physiology, 170(2), 225–229. https://doi.org/10.1016/j.jplph.2012.09.005

El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., Alaux, M., Quesneville, H., Pont, C., & Salse, J. (2017). Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). New Phytologist, 213(3), 1477–1486. https://doi.org/10.1111/nph.14113

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L.,

Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Galasso, I., Piergiovanni, A. R., Lioi, L., Campion, B., Bollini, R., & Sparvoli, F. (2009). Genome organization of Bowman-Birk inhibitor in common bean (*Phaseolus vulgaris* L.). Molecular Breeding, 23(4), 617–624. https://doi.org/10.1007/s11032-009-9260-4

Glover, N. M., Daron, J., Pingault, L., Vandepoele, K., Paux, E., Feuillet, C., & Choulet, F. (2015). Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. Genome Biology, 16(1), 188. https://doi.org/10.1186/s13059-015-0754-6

Habib, H., & Fazili, K. M. (2007). Plant protease inhibitors: A defense strategy in plants. Biotechnology and Molecular Biology Review, 2(3), 68–85.

Haq, S. K., Atif, S. M., & Khan, R. H. (2004). Protein proteinase inhibitor genes in combat against insects, pests, and pathogens: Natural and engineered phytoprotection. Archives of Biochemistry and Biophysics, 431(1), 145–159. https://doi.org/10.1016/j.abb.2004.07.022

Hellinger, R., & Gruber, C. W. (2019). Peptide-based protease inhibitors from plants. Drug Discovery Today, 24(9), 1877–1889. https://doi.org/10.1016/j.drudis.2019.05.026

Hilder, V. A., Gatehouse, A. M. R., Sheerman, S. E., Barker, R. F., & Boulter, D. (1987). A novel mechanism of insect resistance engineered into tobacco. Nature, 330(6144), 160–163. https://doi.org/10.1038/330160a0

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. Molecular Biology and Evolution, 35(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hu, B., Jin, J., Guo, A. Y., Zhang, H., Luo, J., & Gao, G. (2015). GSDS 2.0: An upgraded gene feature visualization server. Bioinformatics, 31(8), 1296–1297. https://doi.org/10.1093/bioinformatics/btu817

James, A. M., Jayasena, A. S., Zhang, J., Berkowitz, O., Secco, D., Knott, G. J., Whelan, J., Bond, C. S., & Mylne, J. S. (2017). Evidence for ancient origins of Bowman-Birk inhibitors from *Selaginella moellendorffii*. Plant Cell, 29(3), 461–473. https://doi.org/10.1105/tpc.16.00831

Jashni, M. K., Mehrabi, R., Collemare, J., Mesarich, C. H., & de Wit, P. J. G. M. (2015). The battle in the apoplast: Further insights into the roles of proteases and their inhibitors in plant–pathogen interactions. In Frontiers in Plant Science (Vol. 6, Issue AUG, p. 584). https://doi.org/10.3389/fpls.2015.00584

Juliana, P., Singh, R. P., Singh, P. K., Poland, J. A., Bergstrom, G. C., Huerta-Espino, J., Bhavani, S., Crossa, J., & Sorrells, M. E. (2018). Genome-wide association mapping for resistance to leaf rust, stripe rust and tan spot in wheat reveals potential candidate genes. Theoretical and Applied Genetics, 131(7), 1405–1422. https://doi.org/10.1007/s00122-018-3086-6

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods, 14(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., Rozewicki, J., & Yamada, K. D. (2018). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics, 20(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Komarnytsky, S., Borisjuk, N., Yakoby, N., Garvey, A., & Raskin, I. (2006). Cosecretion of protease inhibitor stabilizes antibodies produced by plant roots. Plant Physiology, 141(4), 1185–1193. https://doi.org/10.1104/pp.105.074419

Kugler, K. G., Siegwart, G., Nussbaumer, T., Ametz, C., Spannagl, M., Steiner, B., Lemmens, M., Mayer, K. F. X., Buerstmayr, H., & Schweiger, W. (2013). Quantitative trait loci-dependent analysis of a gene co-expression network associated with *Fusarium* head blight resistance in bread wheat (*Triticum aestivum* L.). BMC Genomics, 14(1), 728. https://doi.org/10.1186/1471-2164-14-728

La Rota, M., & Sorrells, M. E. (2004). Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. Functional and Integrative Genomics, 4(1), 34–46. https://doi.org/10.1007/s10142-003-0098-2

Laskowski, M., & Kato, I. (1980). Protein inhibitors of proteinases. Annual Review of Biochemistry, 49(1), 593–626. https://doi.org/10.1146/annurev.bi.49.070180.003113

Laskowski, M., & Qasim, M. A. (2000). What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? Protein Structure and Molecular Enzymology, 1477(1–2), 324–337. https://doi.org/10.1016/S0167-4838(99)00284-8

Lawrence, P. K., & Koundal, K. R. (2002). Plant protease inhibitors in control of phytophagous insects. Electronic Journal of Biotechnology, 5(1), 93–109. https://doi.org/10.2225/vol5-issue1-fulltext-3

Ling, H. Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., Li, Y., Yu, Y., Du, H., Qi, M., Li, Y., Lu, H., Yu, H., Cui, Y., Wang, N., Chen, C., … Liang, C. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature, 557(7705), 424–428. https://doi.org/10.1038/s41586-018-0108-0

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., & Sun, Q. (2015). Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and

drought acclimation in wheat (*Triticum aestivum* L.). BMC Plant Biology, 15(1), 152. https://doi.org/10.1186/s12870-015-0511-8

Lukaszewski, A. J., Alberti, A., Sharpe, A., Kilian, A., Stanca, A. M., Keller, B., Clavijo, B. J., Friebe, B., Gill, B., Wulff, B., Chapman, B., Steuernagel, B., Feuillet, C., Viseux, C., Pozniak, C., Rokhsar, D. S., Klassen, D., Edwards, D., Akhunov, E., … Feuillet, C. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science, 345(6194), 1251788. https://doi.org/10.1126/science.1251788

Luo, M. C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L., Wang, Y., McGuire, P. E., Liu, S., Long, H., Ramasamy, R. K., Rodriguez, J. C., Van Sonny, L., Yuan, L., Wang, Z., Xia, Z., … Dvoák, J. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature, 551(7681), 498–502. https://doi.org/10.1038/nature24486

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. Journal of Genetics, 92(1), 155–161. https://doi.org/10.1007/s12041-013-0212-8

Malefo, M. B., Mathibela, E. O., Crampton, B. G., & Makgopa, M. E. (2020). Investigating the role of Bowman-Birk serine protease inhibitor in *Arabidopsis* plants under drought stress. Plant Physiology and Biochemistry, 149, 286–293. https://doi.org/10.1016/j.plaphy.2020.02.007

Mello, M. O., Tanaka, A. S., & Silva-Filho, M. C. (2003). Molecular evolution of Bowman-Birk type proteinase inhibitors in flowering plants. Molecular Phylogenetics and Evolution, 27(1), 103–112. https://doi.org/10.1016/S1055-7903(02)00373-1

Munkvold, J. D., Greene, R. A., Bermudez-Kandianis, C. E., La Rota, C. M., Edwards, H., Sorrells, S. F., Dake, T., Benscher, D., Kantety, R., Linkiewicz, A. M., Dubcovsky, J., Akhunov, E. D., Dvořák, J., Miftahudin, Gustafson, J. P., Pathan, M. S., Nguyen, H. T., Matthews, D. E., Chao, S., … Sorrells, M. E. (2004). Group 3 chromosome bin maps of wheat and their relationship to rice chromosome 1. Genetics, 168(2), 639–650. https://doi.org/10.1534/genetics.104.034819

Nagasue, A., Fukamachi, H., Ikenaga, H., & Funatsu, G. (1988). The amino acid sequence of barley rootlet trypsin inhibitor. Agricultural and Biological Chemistry, 52(6), 1505–1514. https://doi.org/10.1080/00021369.1988.10868867

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution, 32(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nishino, N., Aoyagi, H., Kato, T., & Izumiya, N. (1977). Studies on the synthesis of proteinase inhibitors: Synthesis and activity of nonapeptide fragments of soybean Bowman-Birk inhibitor. Journal of Biochemistry, 82(3), 901–909. https://doi.org/10.1093/oxfordjournals.jbchem.a131767

Odani, S., Koide, T., & Ono, T. (1986). Wheat germ trypsin inhibitors. Isolation and structural characterization of single-headed and double-headed inhibitors of the Bowman-Birk type. Journal of Biochemistry, 100(4), 975–983. https://doi.org/10.1093/oxfordjournals.jbchem.a121810

Othman, T., Abu Bakar, N., Zainal Abidin, R., Mahmood, M., Saidi, N., & Shaharuddin, N. (2014). Potential of plant's Bowman-Birk protease inhibitor in combating abiotic stresses: A mini review. Bioremediation Science & Technology Research, 2(2), 53–61.

Pak, C., & Van Doorn, W. G. (2005). Delay of Iris flower senescence by protease inhibitors. New Phytologist, 165(2), 473–480. https://doi.org/10.1111/j.1469-8137.2004.01226.x

Pang, Z., Zhou, Z., Yin, D., Lv, Q., Wang, L., Xu, X., Wang, J., Li, X., Zhao, X., Jiang, G., Lan, J., Zhu, L., Hu, S., & Liu, G. (2013). Transgenic rice plants overexpressing *BBTI4* confer partial but broad-spectrum bacterial blight resistance. Journal of Plant Biology, 56(6), 383–390. https://doi.org/10.1007/s12374-013-0277-1

Pearce, S., Vazquez-Gross, H., Herin, S. Y., Hane, D., Wang, Y., Gu, Y. Q., & Dubcovsky, J. (2015). WheatExp: An RNA-seq expression database for polyploid wheat. BMC Plant Biology, 15(1), 299. https://doi.org/10.1186/s12870-015-0692-1

Pekkarinen, A. I., Longstaff, C., & Jones, B. L. (2007). Kinetics of the inhibition of *Fusarium* serine proteinases by barley (*Hordeum vulgare* L.) inhibitors. Journal of Agricultural and Food Chemistry, 55(7), 2736–2742. https://doi.org/10.1021/jf0631777

Phanstiel, D. H., Boyle, A. P., Araya, C. L., & Snyder, M. P. (2014). Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics, 30(19), 2808–2810. https://doi.org/10.1093/bioinformatics/btu379

Poerio, E., Caporale, C., Carrano, L., Caruso, C., Vacca, F., & Buonocore, V. (1994). The amino acid sequence and reactive site of a single-headed trypsin inhibitor from wheat endosperm. Journal of Protein Chemistry, 13(2), 187–194. https://doi.org/10.1007/BF01891977

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. Nucleic Acids Research, 46(W1), W200–W204. https://doi.org/10.1093/nar/gky448

Powell, J. J., Carere, J., Fitzgerald, T. L., Stiller, J., Covarelli, L., Xu, Q., Gubler, F., Colgrave, M. L., Gardiner, D. M., Manners, J. M., Henry, R. J., & Kazan, K. (2017). The *Fusarium* crown

rot pathogen *Fusarium pseudograminearum* triggers a suite of transcriptional and metabolic changes in bread wheat (*Triticum aestivum* L.). Annals of Botany, 119(5), 853–867. https://doi.org/10.1093/aob/mcw207

Prakash, B., Murthy, M. R. N., Sreerama, Y. N., Rao, D. R., & Gowda, L. R. (1997). Studies on simultaneous inhibition of trypsin and chymotrypsin by horsegram Bowman-Birk inhibitor. Journal of Biosciences, 22(5), 545–554. https://doi.org/10.1007/BF02703392

Qi, R. F., Song, Z. W., & Chi, C. W. (2005). Structural features and molecular evolution of Bowman-Birk protease inhibitors and their potential application. Acta Biochimica et Biophysica Sinica, 37(5), 283–292. https://doi.org/10.1111/j.1745-7270.2005.00048.x

Qu, L. J., Chen, J., Liu, M., Pan, N., Okamoto, H., Lin, Z., Li, C., Li, D., Wang, J., Zhu, G., Zhao, X., Chen, X., Gu, H., & Chen, Z. (2003). Molecular cloning and functional analysis of a novel type of Bowman-Birk inhibitor gene family in rice. Plant Physiology, 133(2), 560–570. https://doi.org/10.1104/pp.103.024810

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raj, S. S. S., Kibushi, E., Kurasawa, T., Suzuki, A., Yamane, T., Odani, S., Iwasaki, Y., Yamane, T., & Ashida, T. (2002). Crystal structure of bovine trypsin and wheat germ trypsin inhibitor (I-2b) complex (2:1) at 2.3 Å resolution. Journal of Biochemistry, 132(6), 927–933. https://doi.org/10.1093/oxfordjournals.jbchem.a003306

Ramírez-González, R. H., Borrill, P., Lang, D., Harrington, S. A., Brinton, J., Venturini, L., Davey, M., Jacobs, J., Van Ex, F., Pasha, A., Khedikar, Y., Robinson, S. J., Cory, A. T., Florio, T., Concia, L., Juery, C., Schoonbeek, H., Steuernagel, B., Xiang, D., … Uauy, C. (2018). The

transcriptional landscape of polyploid wheat. Science, 361(6403), eaar6089. https://doi.org/10.1126/science.aar6089

Rawlings, N. D., & Barrett, A. J. (1993). Evolutionary families of peptidases. Biochemical Journal, 290(1), 205–218. https://doi.org/10.1042/bj2900205

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Research, 46(D1), D624–D632. https://doi.org/10.1093/nar/gkx1134

Rawlings, N. D., Tolle, D. P., & Barrett, A. J. (2004). Evolutionary families of peptidase inhibitors. Biochemical Journal, 378(3), 705–716. https://doi.org/10.1042/BJ20031825

Richards, K. D., Snowden, K. C., & Gardner, R. C. (1994). *wali6* and *wali7.* Genes induced by aluminum in wheat (*Triticum aestivum* L.) roots. Plant Physiology, 105(4), 1455–1456. https://doi.org/10.1104/pp.105.4.1455

Rudd, J. J., Kanyuka, K., Hassani-Pak, K., Derbyshire, M., Andongabo, A., Devonshire, J., Lysenko, A., Saqi, M., Desai, N. M., Powers, S. J., Hooper, J., Ambroso, L., Bharti, A., Farmer, A., Hammond-Kosack, K. E., Dietrich, R. A., & Courbot, M. (2015). Transcriptome and metabolite profiling of the infection cycle of *Zymoseptoria tritici* on wheat reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions and a variation on the hemibiotrophic life. Plant Physiology, 167(3), 1158–1185. https://doi.org/10.1104/pp.114.255927

S. Bateman, K., & N.G. James, M. (2011). Plant protein proteinase inhibitors: Structure and mechanism of inhibition. Current Protein & Peptide Science, 12(5), 341–347. https://doi.org/10.2174/138920311796391124

Sari, E., Cabral, A. L., Polley, B., Tan, Y., Hsueh, E., Konkin, D. J., Knox, R. E., Ruan, Y., & Fobert, P. R. (2019). Weighted gene co-expression network analysis unveils gene networks associated with the *Fusarium* head blight resistance in tetraploid wheat. BMC Genom, 20(1), 925. https://doi.org/10.1186/s12864-019-6161-8

Schilling, S., Kennedy, A., Pan, S., Jermiin, L. S., & Melzer, R. (2020). Genome-wide analysis of MIKC-type MADS-box genes in wheat: Pervasive duplications, functional conservation and putative neofunctionalization. New Phytologist, 225(1), 511–529. https://doi.org/10.1111/nph.16122

Schweiger, W., Steiner, B., Vautrin, S., Nussbaumer, T., Siegwart, G., Zamini, M., Jungreithmeier, F., Gratl, V., Lemmens, M., Mayer, K. F. X., Bérgès, H., Adam, G., & Buerstmayr, H. (2016). Suppressed recombination and unique candidate genes in the divergent haplotype encoding *Fhb1*, a major *Fusarium* head blight resistance locus in wheat. Theoretical and Applied Genetics, 129(8), 1607–1623. https://doi.org/10.1007/s00122-016-2727-x

Shan, L., Li, C., Chen, F., Zhao, S., & Xia, G. (2008). A Bowman-Birk type protease inhibitor is involved in the tolerance to salt stress in wheat. Plant, Cell and Environment, 31(8), 1128–1137. https://doi.org/10.1111/j.1365-3040.2008.01825.x

Singh, S., Singh, A., Kumar, S., Mittal, P., & Singh, I. K. (2020). Protease inhibitors: Recent advancement in its usage as a potential biocontrol agent for insect pest management. Insect Science, 27(2), 186–201. https://doi.org/10.1111/1744-7917.12641

Snowden, K. C., Richards, K. D., & Gardner, R. C. (1995). Aluminum-induced genes. Induction by toxic metals, low calcium, and wounding and pattern of expression in root tips. Plant Physiology, 107(2), 341–348. https://doi.org/10.1104/pp.107.2.341

Tan, C. T., Assanga, S., Zhang, G., Rudd, J. C., Haley, S. D., Xue, Q., Ibrahim, A., Bai, G., Zhang, X., Byrne, P., Fuentealba, M. P., & Liu, S. (2017). Development and validation of KASP markers for *wheat streak mosaic virus* resistance gene *Wsm2*. Crop Science, 57(1), 340–349. https://doi.org/10.2135/cropsci2016.04.0234

Tashiro, M., Asao, T., Hirata, C., Takahashi, K., & Kanamori, M. (1990). The complete amino acid sequence of a major trypsin inhibitor from seeds of foxtail millet (*Setaria italica*). Journal of Biochemistry, 108(4), 669–672. https://doi.org/10.1093/oxfordjournals.jbchem.a123260

Tedeschi, F., Di Maro, A., Facchiano, A., Costantini, S., Chambery, A., Bruni, N., Capuzzi, V., Ficca, A. G., & Poerio, E. (2012). Wheat Subtilisin/Chymotrypsin Inhibitor (WSCI) as a scaffold for novel serine protease inhibitors with a given specificity. Molecular BioSystems, 8(12), 3335–3343. https://doi.org/10.1039/c2mb25320h

Uauy, C., Wulff, B. B. H., & Dubcovsky, J. (2017). Combining traditional mutagenesis with new high-throughput sequencing and genome editing to reveal hidden variation in polyploid wheat. Annual Review of Genetics, 51(1), 435–454. https://doi.org/10.1146/annurev-genet-120116-024533

USDA-FAO. (2018). Food and Agriculture Organization of the United Nations Database. https://doi.org/10.1016/B978-0-12-384947-2.00270-1

Volpicella, M., Leoni, C., Costanza, A., De Leo, F., Gallerani, R., & R. Ceci, L. (2011). Cystatins, serpins and other families of protease inhibitors in plants. In Current Protein & Peptide Science (Vol. 12, Issue 5, pp. 386–398). https://doi.org/10.2174/138920311796391098

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D., … Pozniak, C. J.

(2020). Multiple wheat genomes reveal global variation in modern breeding. Nature, 588(7837), 277–283. https://doi.org/10.1038/s41586-020-2961-x

Woodcroft, B. J., Boyd, J. A., & Tyson, G. W. (2016). OrfM: a fast open reading frame predictor for metagenomic data. Bioinformatics, 32(17), 2702–2703. https://doi.org/10.1093/bioinformatics/btw241

Xu, D., Xue, Q., McElroy, D., Mawal, Y., Hilder, V. A., & Wu, R. (1996). Constitutive expression of a cowpea trypsin inhibitor gene, *CpTi*, in transgenic rice plants confers resistance to two major rice insect pests. Molecular Breeding, 2(2), 167–173. https://doi.org/10.1007/BF00441431

Yan, K. M., Chang, T., Soon, S. A., & Huang, F. Y. (2009). Purification and characterization of Bowman-Birk protease inhibitor from rice coleoptiles. Journal of the Chinese Chemical Society, 56(5), 949–960. https://doi.org/10.1002/jccs.200900139

Yang, F., Li, W., & Jørgensen, H. J. L. (2013). Transcriptional reprogramming of wheat and the hemibiotrophic pathogen *Septoria tritici* during two phases of the compatible interaction. PLoS ONE, 8(11), 81606. https://doi.org/10.1371/journal.pone.0081606

Ye, X. Y., Ng, T. B., & Rao, P. F. (2001). A Bowman-Birk-type trypsin-chymotrypsin inhibitor from broad beans. Biochemical and Biophysical Research Communications, 289(1), 91–96. https://doi.org/10.1006/bbrc.2001.5965

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution, 8(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Yu, Y., Ouyang, Y., & Yao, W. (2018). ShinyCircos: An R/Shiny application for interactive creation of Circos plot. Bioinformatics, 34(7), 1229–1231. https://doi.org/10.1093/bioinformatics/btx763

Zhang, C., Fang, H., Shi, X., He, F., Wang, R., Fan, J., Bai, P., Wang, J., Park, C. H., Bellizzi, M., Zhou, X., Wang, G. L., & Ning, Y. (2020). A fungal effector and a rice NLR protein have antagonistic effects on a Bowman–Birk trypsin inhibitor. Plant Biotechnology Journal, n/a(n/a). https://doi.org/10.1111/pbi.13400

Zhang, H., Yang, Y., Wang, C., Liu, M., Li, H., Fu, Y., Wang, Y., Nie, Y., Liu, X., & Ji, W. (2014). Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew. BMC Genom, 15(1), 898. https://doi.org/10.1186/1471-2164-15-898

Zhang, L., Nakanishi Itai, R., Yamakawa, T., Nakanishi, H., Nishizawa, N. K., & Kobayashi, T. (2014). The Bowman-Birk trypsin inhibitor IBP1 interacts with and prevents degradation of IDEF1 in rice. Plant Molecular Biology Reporter, 32(4), 841–851. https://doi.org/10.1007/s11105-013-0695-8

Zhang, Z., Zhang, Z., Gou, X., Gou, X., Xun, H., Xun, H., Bian, Y., Bian, Y., Ma, X., Li, J., Li, N., Gong, L., Feldman, M., Liu, B., & Levy, A. A. (2020). Homoeologous exchanges occur through intragenic recombination generating novel transcripts and proteins in wheat and other polyploids. Proceedings of the National Academy of Sciences of the United States of America, 117(25), 14561–14571. https://doi.org/10.1073/pnas.2003505117

CHAPTER 3. GENOMIC AND MOLECULAR CHARACTERIZATION OF THE *WHEAT STREAK MOSAIC VIRUS* RESISTANCE LOCUS 2 (*WSM2*) IN COMMON WHEAT (*TRITICUM AESTIVUM*. L)

3.1 Summary

WSMV is an economically important viral pathogen that negatively impacts global wheat production, particularly in the Great Plains region of the U.S. The WSMV resistance locus *Wsm2* provides strong resistance and has been widely deployed, but the underlying causative genes has not been cloned, limiting our ability to understand the resistance mechanisms and to develop perfect markers for breeding. In this study, the *Wsm2* interval was analyzed using wheat reference genomes, from which 94 candidate genes were identified. Haplotype analysis of *Wsm2* in seventeen wheat cultivars collected from different agroecological zones indicated that the *Wsm2* locus is in a dynamic region of the genome and is a rare allele likely absent from many modern wheat cultivars. Examination of the *Wsm2* locus using exome reads from 'Snowmass', a cultivar carrying the *Wsm2* locus, identified gene copy number duplications for four adjacent UDP-glycotransferase genes and other natural genetic variation underlying *Wsm2*. Through *de novo* assembly of RNA-seq reads that do not map to the wheat reference genome 'Chinese Spring', three unique transcripts specific to lines that contain *Wsm2* locus (*Wsm2*+) were identified. Furthermore, five candidate genes within the *Wsm2* interval were differentially expressed between *Wsm2*+ and lines absent for *Wsm2* locus (*Wsm2*-) following WSMV infection and their expression in RNA-seq were validated with qRT-PCR, these candidates are possible causative genes underlying *Wsm2*. For one of the candidates, annotated as *RPM1*, CRISPR/Cas9-edited plants carrying gene knockouts were developed for functional characterization of its molecular function in WSMV resistance.

3.2 Introduction

Common wheat (*Triticum aestivum* L.) is one of the most important crops worldwide, providing approximately 20% of the calories and proteins consumed by the human population (FAOSTAT, 2020). *Wheat streak mosaic virus* (WSMV) is an economically important viral pathogen that threatens wheat production around the globe (Navia et al., 2013). In the United States, WSMV mainly affects crops grown in the Great Plains region, causing average annual yield losses of approximately 5%, although severe localized infections can result in complete crop failure (Singh & Kundu, 2018; McKelvy et al., 2021). Once infected with WSMV, wheat leaves exhibit a characteristic yellow and green streaked mosaic pattern (Hadi et al., 2011). For winter wheat cultivars, symptoms are most severe when infection occurs during tillering, and can include stunting, poor fertility, and reduced grain set (Hunger et al., 1992).

WSMV is the type species of the genus *Tritimovirus* within the family *Potyviridae* (Stenger et al., 1998). The WSMV genome consists of a single positive-strand genomic RNA of 9,384 nucleotides (nt) encoding a polyprotein of ~ 350 kDa, which is processed into ten mature proteins after cleavage by three proteinases encoded in the viral genome (P1, HC-Pro and NIa-Pro) (Tatineni & French, 2014). To systemically infect hosts, plant viruses require different viral components to establish initial local infection, followed by cell-to-cell movement through plasmodesmata and long-distance movement through the vascular tissue (Seo & Kim, 2016). The WSMV coat protein (CP) is required for its transmission by vectors (Tatineni & Hein, 2018) but can tolerate extensive point mutations and small insertions and deletions (Indels) while still retaining the ability to systemically infect wheat (Tatineni et al., 2014; Tatineni & French, 2014).

The only known transmission vector for WSMV is the eriophyid wheat curl mite (WCM), *Aceria tosichella* Keifer, which has a body length of ~200 μm and is spread between crops by the

wind (Slykhuis, 1955). WCM is also the transmission vector for *Triticum mosaic virus* (TriMV) (Chuang et al., 2017), which interacts synergistically with WSMV in the field (Seifers et al., 2009; Tatineni et al., 2019). WSMV can be picked up by WCM from infected host plants during a 10- to 30-minute feeding time and remains active in WCM for 7-9 days (Singh & Kundu, 2018). However, the WSMV carried by WCM cannot be passed to the next WCM generation through the egg stage (Singh & Kundu, 2018). Upon landing on wheat plants, the WCM remains hidden in rolled and curled leaves as well as leaf sheaths of wheat seedlings, and can survive for several months (Navia et al., 2013). As a result, there are no effective miticides against WCM (Navia et al., 2013). Moreover, growing plants in the field such as volunteer wheat, and other monocots and wild weeds, including oats (*Avena sativa*), barley (*Hordeum vulgare*), rye (*Secale cereale*), corn (*Zea mays*) and foxtail millet (*Setaria italica*), can serve as a 'green bridge' for WCM to complete their life cycle between wheat cropping seasons (Singh & Kundu, 2018). This broad range of hosts makes it ineffective and impractical for many growers to use cultural practices to completely eradicate WCM from infected fields (Singh et al., 2018). Therefore, the long-term strategy to prevent damage caused by WCM and WSMV is to develop wheat cultivars with genetic resistance to the WSMV-WCM disease complex (Harvey et al., 1999).

To date, four quantitative trait loci (QTL) associated with WCM resistance have been identified from grass species and ancestors of common wheat (Thomas & Conner, 1986; Whelan & Hart, 1988; Malik et al., 2003). Although these resistance alleles inhibit WCM reproductive potential and reduce its transmission rate in the field, the effectiveness varies between WCM populations (Murugan et al., 2011). Moreover, because these loci are all derived from alien introgressions, they are associated with linkage drag and the host genetic resistance to WCM varies according to different field conditions (Harvey et al., 1999).

In addition to introgressing genetic resistance to WCM, efforts have been made to identify alleles with direct resistance to WSMV itself. Four QTLs (*Wsm1, Wsm2*, *Wsm3* and *c2652*) for WSMV resistance have been identified to date (Haley et al., 2002; Sharp et al., 2002; Haber et al., 2006; Divis et al., 2006). Both *Wsm1* and *Wsm3* also confer resistance to TriMV (Seifers et al., 2009). *Wsm1* and *Wsm3* originated in intermediate wheatgrass (*Thinopyrum intermedium*) and were transferred into common wheat cultivars through alien translocation and have only been mapped at low resolution to a whole chromosome arm (Wells et al., 1982; Friebe et al., 2009; W. Liu et al., 2011; Danilova et al., 2017). When deployed in elite cultivars, these alleles confer a yield penalty due to linkage drag, limiting their value in wheat breeding program. For example, in the absence of WSMV infection, *Wsm1* confers a yield penalty of up to 30% (Seifers et al., 1995; Sharp et al., 2002). *c2652* was identified from a hard red spring wheat population with an unknown pedigree (Haber et al., 2006) and has not been utilized in wheat germplasm development. In contrast, *Wsm2* was first identified from a wheat breeding line CO960293-2 and most likely originated in a common wheat background (Haley et al., 2002). Among the four QTLs conferring resistance to WSMV, *Wsm2* is the only locus that has been mapped to a relatively short genomic region (Zhang & Hua, 2018).

Despite one recently discovered WSMV isolate from *Setaria viridis* that can break *Wsm2*-mediated resistance (Kumssa et al., 2019), *Wsm2* provides strong resistance to a variety of WSMV strains (Seifers et al., 2006; Lu et al., 2012). The *Wsm2*-mediated WSMV resistance was effective and led to consistently low WSMV incidence in field conditions, highlighting the value of deploying *Wsm2* in breeding programs to control WSMV (McKelvy et al., 2021). Moreover, there is no evidence of deleterious impacts on yield and agronomic traits associated with *Wsm2* (Lu et al., 2012). *Wsm2* is temperature sensitive and is less effective in field temperatures above 18 °C

(Seifers et al., 2006). *Wsm2* has been introduced into several common wheat cultivars by recombination, including 'RonL' (J. T. Martin et al., 2007), 'Snowmass' (Haley et al., 2011), 'Clara CL' (Martin et al., 2014), 'Oakley CL' (Zhang et al., 2015), and 'Joe' (Zhang et al., 2016).

Linkage mapping in two $F_{2:3}$ populations showed that the inheritance of the WSMV resistance underlying *Wsm2* is controlled by a single dominant allele located on chromosome arm 3BS (Lu et al., 2011). Subsequent studies further mapped the *Wsm2* locus to a region of less than 1 cM in a recombinant inbred line (RIL) population (Assanga et al., 2017; Tan et al., 2017). Four SNP markers tightly linked to *Wsm2* were transformed into KASP assays and validated in a RIL population ('CO960293' × 'TAM111') and in two doubled haploid populations ('RonL' × 'Ripper' and 'Snowmass' × 'Antero'), from which haplotypes associated with WSMV resistance and susceptibility, respectively, were identified (Tan et al., 2017). Two of the tightly linked SNP markers located at 16.4 Mbp, one at 17.8 Mbp, and a left boundary flanking marker defined by recombination in the RIL population was mapped to 18.9 Mbp physical position on the wheat reference genome (Tan et al., 2017). A genome wide association study (GWAS) performed on 597 wheat breeding lines identified ten other significant SNP markers associated with WSMV resistance (Dhakal et al., 2018). These ten SNPs mapped to a physical interval flanking a 17.1 – 18.9 Mbp telomeric region on chromosome 3B in the wheat reference genome IWGSC RefSeq v1.0, coinciding with the *Wsm2* locus. This region contains a high density of genes encoding Bowman-Birk inhibitors (BBI) and is highly diverse between wheat cultivars with evidence of structural variation (Xie et al., 2021).

Fewer resistance (*R*) genes have been characterized for viral pathogens than for either bacterial or fungal pathogens (Kourelis & van der Hoorn, 2018), partly due to our more limited understanding of plant immunity mechanisms for plant viruses (Ronde et al., 2014). The majority

of cloned single dominant antiviral *R* genes to date belongs to the nucleotide binding site leucine-rich repeat (NB-LRR) type, that recognize avirulence factors (Avr) encoded by viral pathogens to trigger Effector-Triggered Immunity (ETI), which is generally associated with hypersensitive response (HR) or the production of salicylic acid (SA) (Ronde et al., 2014). Some dominant *R* genes against viruses have different resistance mechanisms from the classical NB-LRR type *R* gene and do not induce HR or SA production (Cosson et al., 2012). For example, *RTM1*, *RTM2*, and *RTM3* cloned in *Arabidopsis thaliana* encode lectin proteins that restrict the long-distance movement of *Tobacco Etch Virus* (TEV) (Whitham et al., 2000; Chisholm et al., 2001; Cosson et al., 2010). Another example is the *Tm-1* gene cloned in tomato that encodes a protein containing a TIM-barrel domain that inhibits *Tomato mosaic virus* (TMV) replication (Ishibashi et al., 2007; Ishibashi & Ishikawa, 2013). Similarly, *Wsm2*-mediated resistance against WSMV does not induce HR and confers resistance to WSMV by impeding long-distance viral movement (Tatineni et al., 2016).

Despite these studies that shed some light on *Wsm2*-mediated resistance, the causative genes underlying *Wsm2*, or any other *WSMV* resistance gene, has yet to be cloned. Identifying the causative gene underlying *Wsm2* will allow for experiments to characterize the molecular role of *Wsm2*, look for alternative allelic variation conferring resistance to novel WSMV isolates, and to characterize the functional pathways by which WSMV resistance can be induced (Skoracka et al., 2018). With better understanding and knowledge of the viral resistance genes, further studies can apply genome editing tools for targeted manipulation of resistance genes to help breed WSMV-resistant wheat cultivars more efficiently.

Recent advances in genomic resources have facilitated genetic mapping and functional genomics studies in wheat. The wheat reference genome assembly has been annotated for the

landrace ‘Chinese Spring’ (Appels et al., 2018; Zhu et al., 2021). Beyond this reference sequence, genome assemblies of sixteen other wheat cultivars facilitate studies of within-species genomic diversity (Walkowiak et al., 2020). Whole genome sequencing (WGS) and assembly remains expensive and technically challenging in common wheat due to its large, polyploid genome (Lukaszewski et al., 2014). Instead, exome capture sequencing or transcriptome sequencing can help reduce sequencing costs while providing informative genetic variation in individual advanced lines and to understand the molecular mechanisms underlying key aspects of plant development and defense responses.

In this study, the *Wsm2* locus was found to lie in a dynamic region of the wheat genome characterized by structural variation among different wheat cultivars. Genotypic and phenotypic data suggest the *Wsm2* gene is rare among modern wheat cultivars, but its presence was confirmed in a ‘Snowmass’ mapping population. Exome capture reads from ‘Snowmass’ revealed two high impact genetic variants together with seven CNVs underlying *Wsm2*. Additionally, an RNA-seq study was performed on ‘Snowmass’ derived doubled haploid lines that also carry *Wsm2* locus (*Wsm2*+) and the *de novo* assembly of the unmapped transcriptomic reads identified three unique transcripts absent from the wheat reference genome. Additionally, five genes within the *Wsm2* locus were found differentially expressed between genotypes, including a gene annotated as *RPM1* that contains an LRR domain. Functional validation of *RPM1* was performed using CRISPR/Cas9 to generate wheat knockout mutants, which will be characterized using WSMV inoculation assays.

3.3 Materials and Methods

3.3.1 Plant materials

Seeds of the wheat cultivars ‘Jagger’, ‘SY Mattis’, ‘Robigus’, ‘Mace’, ‘Paragon’, ‘Landmark’, ‘Stanley’, ‘Claire’, ‘Weebill’, ‘Cadenza’, ‘Kronos’, and ‘Chinese Spring’ were obtained from

Seedstor (https://www.seedstor.ac.uk/search-browseaccessions.php?idCollection=35) and were used to perform WSMV phenotyping assays. A *Triticum aestivum* L. doubled haploid (DH) population (n = 116) produced by Heartland Plant Innovations Inc. was developed by wheat-maize wide hybridization (Santra et al., 2017) from the parents 'Snowmass' (WSMV resistant) and 'Antero' (WSMV susceptible) and used for linkage mapping. 'Snowmass' and 'Antero' leaf tissues were used to quantify WSMV by qRT-PCR in a time course from 0, 5, 10, and 15-day post inoculation (dpi). For the RNA-seq study, eight individuals homozygous for the *Wsm2* locus were selected from the DH population as plant materials. Using GBS markers within 16.5 Mbp to 18.8 Mbp on IWGSC RefSeq v1.0 genome, four individuals were shown to have an identical haplotype to 'Snowmass' (*Wsm2*+), and another four had an identical haplotype to 'Antero' (*Wsm2*-) (Table S3.1). Meanwhile, the phenotype scores confirmed that *Wsm2*+ DH lines are resistant to WSMV, whereas *Wsm2*- DH lines are susceptible (Table S3.1). Additionally, the exome reads of 'Snowmass', 'Antero', 'Brawl', 'Byrd', 'Hatcher', 'C0940610', and 'Platte' were captured using the NimbleGen SeqCap EZ wheat whole-genome assay and sequenced as described in Jordan et al., 2015 and He et al., 2019.

3.3.2 WSMV inoculation and phenotype evaluation

An isolate of WSMV originally collected from Akron, Colorado in 2017 was propagated in the greenhouse by mechanically inoculating on susceptible winter wheat genotype 'Longhorn' every six months. Leaf tissues with a yellow streaking or mosaic pattern typical of WSMV were collected, frozen at -80 °C and used to prepare fresh inoculum. The inoculum was prepared with 1:10 (w/v) dilution of the WSMV-infected wheat leaf tissue and 0.01 M potassium phosphate buffer (pH 7.4) and inoculated on two-week-old seedlings. To determine the effectiveness of the *Wsm2* allele, wheat plants were grown in a PGR15 growth chamber (Conviron, Manitoba, Canada)

in a 12 h photoperiod with temperatures set to 18 ℃ day/15 ℃ night. Mechanical inoculation was performed using a soft sponge soaked with inoculum that was gently rubbed on the surface of wheat seedling leaves that were previously dusted with Carborundum powder. Phenotyping was evaluated two and three weeks after inoculation of WSMV by examination of visual symptoms based on a 1-5 scale; (1 = no chlorosis; 2 = a few chlorotic streaks; 3 = moderate mosaic; 4 = severe mosaic; 5 = severe mosaic, necrosis, and yellowing) (Tan et al., 2017). Plants with scores ≤ 2 were considered resistant, and plants with scores > 2 were considered susceptible.

3.3.3 Haplotype analysis for within-species variation at the *Wsm2* locus

BLAST alignment of 100 bp surrounding sequences (Total 201 bp sequence with the SNP at position 101 bp) of four tightly linked SNP markers (IAAV6442, BS00018764_51, IWA7647 and BS00026471_51, Table S3.2) identified the physical position of *Wsm2* on the wheat reference genomes IWGSC RefSeq v1.0 (Appels et al., 2018) and RefSeq v2.1 (Zhu et al., 2021). These SNPs were also mapped to the genome assemblies of sixteen other wheat cultivars (Mace, Lancer, CDC Stanley, CDC Landmark, Julius, Norin61, ArinaLrFor, Jagger, Cadenza, Paragon, Kronos, Robigus, Claire, Spelt, Weebill, SY Mattis) to analyze genomic variation using the Galaxy platform (Afgan et al., 2018).

3.3.4 Linkage mapping analysis for the DH population

The DH population was subjected to genotyping-by-sequencing (GBS, Elshire et al., 2011) and data was processed as described in Liu et al., (2016). The GBS markers were mapped to wheat reference genome IWGSC RefSeq v1.0 (Appels et al., 2018) and annotated for their position, for example, S3B_16589830 indicates the marker is placed on wheat chromosome 3B at the physical position 16,589,830 bp. Quantitative trait loci (QTL) analysis was performed with R version 4.0.3

package qtl (Arends et al., 2010) and ASMap (Taylor & Butler, 2017) considering the mean phenotyping scores from four biological replicates of each DH line.

3.3.5 Exome capture analysis for genomic variations underlying *Wsm2* locus

Raw paired-end Illumina reads were filtered for quality using fastp (Chen et al., 2018). Reads were then aligned to the wheat reference IWGSC RefSeq v1.0 assembly using bowtie2 v. 2.3.5 (Langmead & Salzberg, 2012) with the following parameters: -k 2 -N 1 -L 22 -D 20 -R 3. The alignments were subjected to samtools v1.11 to generate sorted BAM files and then bcftools v1.11 (Danecek et al., 2021) was used to and call variants within *Wsm2* locus with mpileup command. The SnpEff (Cingolani et al., 2012) tool was used to predict the effects of genetic variants, including SNPs, indels, and multiple-nucleotide polymorphisms. The sorted BAM files from 'Antero', 'Brawl', 'Byrd', 'Hatcher', 'C0940610', and 'Platte' were used as a reference set to assay copy number variation (CNV) for the test sample 'Snowmass' (*Wsm2*+) using the ExomeDepth R package (Plagnol et al., 2012). The default parameters were used, except for transition probability = 0.001 ("CallCNVs") and min.overlap = 0.01 ("AnnotateExtra").

3.3.6 WSMV quantification

The amount of WSMV cDNA in whole leaf tissues from 'Antero' and 'Snowmass' was quantified with qRT-PCR in five biological replications over the time course of 0, 5, 10, and 15 days after inoculation with WSMV (dpi). Total RNA from leaf samples were isolated with spectrum total RNA kit (Sigma, USA), followed by on-column DNase I digestion treatment (Sigma-Aldrich) to remove genomic DNA. WSMV was detected via qRT-PCR using coat protein (CP) specific primers and the probe listed in Table S3.3. The one-step RT-PCR reaction was carried out in the Taqman master mix reagent kits (Applied biosystems) to quantify the absolute amount of WSMV.

3.3.7 RNA-seq library preparation

Sixteen samples were collected for the RNA-seq experiment, including four biological replicates of two genotypes (*Wsm2*+ and *Wsm2*-) and two treatments (mock inoculation with phosphate buffer (C) and WSMV inoculation with infected tissue (T)). RNA-seq samples were labeled as *Wsm2*- (C), *Wsm2*- (T), *Wsm2*+ (C) and *Wsm2*+ (T). Whole leaf tissue was harvested at 10 dpi, stored at -80 ℃ and ground to a homogenized fine powder in liquid nitrogen. Total RNA was isolated with spectrum total RNA kit (Sigma, USA) and quantified using Qubit. The Agilent 2100 bioanalyzer (RNA Nano Chip, Agilent, CA) was used to check RNA integrity. The library construction and sequencing via Illumina HiSeq 2000 were performed by Novogene Co., Ltd (Sacramento, CA, USA), approximately 150 bp paired end (PE) raw reads were generated.

3.3.8 Transcript abundance, differential expression (DE) and GO enrichment analysis

To quantify WSMV reads in the RNA-seq samples, the coding sequence from WSMV isolate KSHm2014 (9,384 bp) was retrieved from NCBI database (MK318278.1, https://www.ncbi.nlm.nih.gov/). This sequence was concatenated to the IWGSC RefSeq v1.0 wheat genome (Appels et al., 2018) as an additional FASTA entry, and used as the reference genome to build index files for alignments. Raw reads of each paired-end library were examined for sequence quality and adaptor sequences were removed using Fastp with default settings (Chen et al., 2018). Trimmed paired-end RNA-seq reads were aligned to the reference genome using STAR (Dobin & Gingeras, 2015) with parameters "-outFilterMismatchNmax 6 -alignIntronMax 10000". Non-normalized reads were counted with featureCounts (Liao et al., 2014) with parameters "-t gene -p" and used as input for the R package "DEseq2" (Love et al., 2014). Read counts of WSMV in each sample were normalized to clean reads of the corresponding samples for count per million (cpm) of WSMV. Differentially expressed genes (DEGs) were identified from

pairwise comparisons for treatment effect (WSMV-treated *vs.* mock-treated), and for genotypic effect under each condition: Resistant *vs.* Susceptible under WSMV-treated condition (T_{Wsm2-}^{Wsm2+}) and under mock-treated condition (C_{Wsm2-}^{Wsm2+}). The *P*-value threshold was determined using Benjamini and Hochberg's approach (Benjamini & Hochberg, 1995) for controlling the false discovery rate (FDR < 0.01) without controlling the $\log_2$ fold change (FC). Venn diagrams were drawn using VENNY software (Oliveros, 2007). Gene ontology (GO) enrichment analyses were performed with the TopGO R package (Alexa A, 2021) and Fisher tests were conducted to identify significant GO terms ($P < 0.01$).

3.3.9 *De novo* transcriptome assembly of unmapped reads in *Wsm2*+ and presence/absence analysis of unique transcripts in the wheat pangenomes

During the STAR alignment, reads that did not map to IWGSC RefSeq v1.0 were collected with parameter "-outReadsUnmapped" and assembled with Trinity (Grabherr et al., 2011) using the following parameters: "-seqType fq –samples_file<input_file> -max_memory 10G -CPU 20". The proportion of reads mapped to the assembly was assessed with Bowtie2 (Langmead & Salzberg, 2012). Then CD-HIT (cd-hit-est – c 0.95) was used to remove redundant transcripts. The assembled transcripts were used as BLASTn queries against the NCBI nucleotide database (cutoff: 1e-5) to annotate gene function. DEG analysis were performed on unmapped reads against assembled transcriptomes to identify candidate genes overlapped between the comparison of T_{Wsm2-}^{Wsm2+} and $Wsm2 +_{C}^{T}$. The transcript sequences for overlapped DEG candidates that considered as unique transcripts in *Wsm2*+ compared to 'Chinese Spring' were extracted and used as queries in BLASTn searches (default: 1e-3) against the coding sequences (CDS) of ten wheat cultivars using the Galaxy platform (Afgan et al., 2018) for presence and absence (PAV) analysis. The PAV analysis for unique candidates among wheat cultivars was performed using the criteria that:

presence (+) was indicated as the top BLAST hit having a CDS similarity percentage > 96%. Absence (-) was indicated when no BLAST output was returned or when the top BLAST hit had a CDS similarity percentage < 96%. For unique transcripts in *Wsm2*+ that are also present in other wheat cultivars, the corresponding gene ID and physical position were extracted using the Galaxy platform (Afgan et al., 2018).

3.3.10 Gene expression validation with qRT-PCR

Transcript levels of selected candidate DEGs identified from RNA-seq experiments were validated with qRT-PCR using the same samples used for WSMV quantification. First-strand cDNA was synthesized from 2 µg of total RNA using SuperScript IV Reverse Transcriptase Kit (Thermo Fisher Scientific, USA) according to the manufacturer's instructions. The qRT-PCR were performed using PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, USA) in a 20 µL reaction with 100 ng cDNA and 1 µL of a 10 µM solution for each primer. Relative gene expression analysis was calculated using *ACTIN* as the internal control gene and $2^{-\Delta CT}$ method was used for relative quantification. Primer efficiency and specificity were determined by analyzing amplification in a four-fold dilution series and checking the dissociation curve for a single amplified product and calculated as: Efficiency (%) = $(4^{\frac{1}{slope}} - 1) \times 100$. All primers in this study had an efficiency greater than 90% and are listed in Table S3.3.

3.3.11 Functional validation of *TraesCS3B02G035800* with CRISPR/Cas9

To characterize the function of *TraesCS3B02G035800*, CRISPR/Cas9 was used to generate gene knockout mutants in elite wheat cultivars. The coding sequence was used as a query in a BLAST search to identify homoeologous copies. Two sgRNAs which are 500 bp apart and targeted at the first exon of this gene (Figure S3.1) were designed based on WheatCRISPR (Cram et al., 2019) web tool by inputting candidate gene name and selecting for targeting at coding sequences

for all homoeologous copies (Table S3.4). Primers to clone sgRNAs are listed in Table S3.4. To clone sgRNA, the JD633-*GRF4-GIF1/CRISPR-Cas9* vector (Debernardi et al., 2020) was used and sgRNA sequences were introduced by GoldenGate reaction into two *Aar*I sites of the vector and transformed into chemically competent *Escherichia coli* DH5α cells. The JD633-*GRF4-GIF1/CRISPR-Cas9-gRNA* vector was validated by Sanger sequencing using primer pairs TaU6-promoter/gRNA-scaffold (Table S3.3). The constructs were first transformed into *Agrobacterium* strain AGL1 by electroporation and then into 'Snowmass' and 'Snowmass 2.0' wheat embryos with *Agrobacterium*-mediated transformation (Hayta et al., 2019). Genomic DNA was extracted from leaf tissues using a standard CTAB DNA extraction method and a PCR assay was performed to detect the presence of Cas9 in the DNA samples (initial denaturation at 95 ℃ for 5 min; 40 cycles of 95℃ for 30 s, 60 ℃ for 1 min, and 68 ℃ for 1 min; and a final extension at 68 ℃ for 5 min). The genotyping assay primers (Table S3.3 and Figure S3.1) were designed using Primer3 (Koressaar & Remm, 2007) and amplification products were then subjected to Sanger sequencing with integrated DNA technology (IDT) company.

3.4 Results

3.4.1 Genomic characterization of *Wsm2* in common wheat

3.4.1.1 Analysis of within-species genomic variation for Wsm2 in seventeen wheat cultivars

To refine the physical position of the *Wsm2* locus, four SNP markers each within 1 cM of *Wsm2* from previous mapping experiments (Table S3.2) were mapped to a 2.4 Mbp region (16.5 Mbp to 18.9 Mbp) on chromosome arm 3BS of the IWGSC RefSeq v1.0 genome assembly (Table S3.5). In the IWGSC RefSeqv2.1 assembly, these markers spanned a 2.5 Mbp region on the same chromosome arm (22.0 Mbp to 24.5 Mbp, Table S3.5) due to a 100 kbp insertion at approximately 24.2 Mbp (Figure S3.2). In both references, this region included the same 94 annotated gene

models (50 high confidence, 44 low confidence, Table S3.6). The wheat IWGSC RefSeq v1.0 reference genome is based on the landrace 'Chinese Spring', which carries the WSMV-susceptibility haplotype across the four SNP markers and exhibits a susceptible phenotype with a mean score equal to 4.0 two weeks post inoculation (Figure 3.1). The results show that the wheat reference genome IWGSC RefSeq v1.0 likely does not contain the *Wsm2* genetic variation.

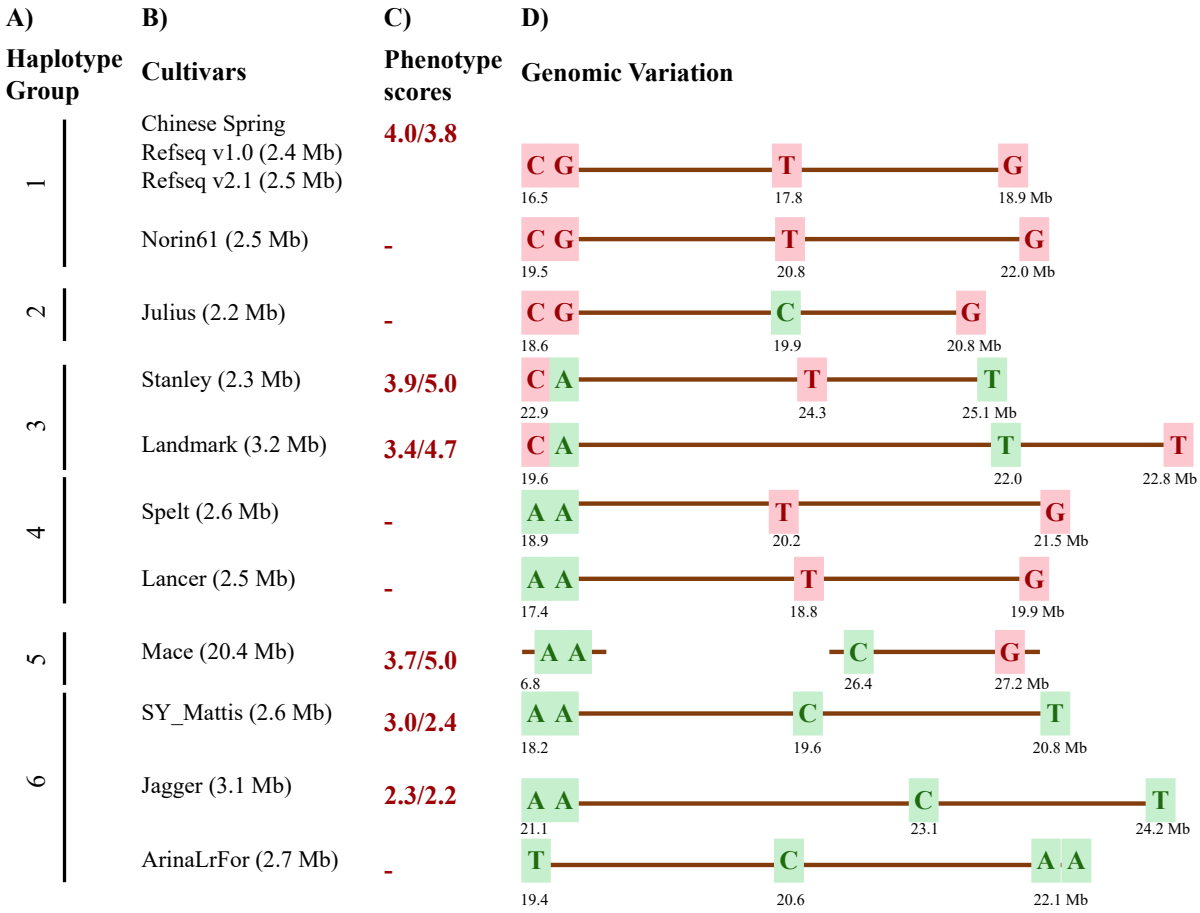


Figure 3.1. Haplotype and genomic variation underlying *Wsm2* in eleven wheat cultivars. **A)** Haplotypes were grouped based on allele type. **B)**. Cultivar names and the *Wsm2* interval size. **C)** Phenotype scores were measured two- and three-week post WSMV inoculation, scores separated by slash. Phenotype scores for seeds not available for subjecting to WSMV screening were indicated dash **D)** The resistant haplotype allele type was highlighted with green, whereas susceptible allele type in pink. Scores underlying the allele type indicate the physical position (Mbp) in each respective genome assembly.

To compare within-species genetic diversity at the *Wsm2* locus, the corresponding genomic region of each SNP marker was mapped to the physical position in ten additional wheat cultivars with pseudomolecule genome assemblies (Figure 3.1). This region contained large structural variation between cultivars. For example, the physical distance between four markers spans from 2.2 Mbp in 'Julius' up to 20.4 Mb in 'Mace' due to a 19.6 Mbp insertion between KASP2 and KASP3 markers (Figure 3.1). Moreover, the order of the four markers was inverted in 'ArinaLrFor', suggesting genomic inversions of *Wsm2* locus in this cultivar.

To further analyze *Wsm2* haplotypes, another six wheat cultivars with scaffold-level genome assemblies were included in the analysis. In total, eight haplotypes in this region were identified among all seventeen wheat cultivars (Table 3.1). The 'CGTG' haplotype associated with WSMV susceptibility was identified in 'Chinese Spring' and 'Norin61', whereas the resistance haplotype 'AACT' was identified in 'Robigus', 'SY Mattis', 'Jagger', and 'ArinaLrFor' (Table 3.1). Moreover, a combination of another six haplotypes for these four markers were identified in the remaining twelve wheat cultivars. However, the WSMV phenotyping screening assay showed these wheat cultivars, regardless of their haplotypes, were all susceptible to WSMV infection (score > 2) two or three weeks after inoculation (Table 3.1). The lack of association between WSMV resistance and haplotypes suggested that these markers are not predictive to identify *Wsm2* when tested in wheat cultivars with quite diverse genetic background. Altogether, these results suggested *Wsm2* lies in a highly dynamic region of wheat genome and is likely absent from all wheat cultivars with assembled genomes, making it a rare allele that likely just present in a few wheat cultivars that developed recently during wheat evolution.

Table 3.1. Haplotype analysis for *Wsm2* in seventeen wheat cultivars. The resistant haplotype alleles or phenotype were indicated as green, whereas susceptible haplotype alleles or phenotype were indicated as pink.

Marker Name	**KASP1**	**KASP2**	**KASP3**	**KASP4**	**Phenotyping**	
Marker Position (bp)	**16,443,466**	**16,455,416**	**17,770,942**	**18,873,524**	**14 dpi**	**21 dpi**
Susceptible	**C**	**G**	**T**	**G**	3.3	3.9
Chinese Spring	C	G	T	G	4.0	3.8
Norin61	C	G	T	G	a	a
Julius	C	G	C	G	a	a
Kronos	C	A	C	G	4.0	4.6
Cadenza	C	A	T	G	3.2	3.0
Weebill	C	A	T	T	3.3	4.3
Clarie	C	A	T	T	2.8	3.3
CDC Stanley	C	A	T	T	3.9	5.0
CDC Landmark	C	A	T	T	3.4	4.7
Spelt	A	A	T	G	a	a
Lancer	A	A	T	G	a	a
Paragon	A	A	C	G	2.1	2.1
Mace	A	A	C	G	3.7	5.0
Robigus	A	A	C	T	3.4	4.6
SY_Mattis	A	A	C	T	3.0	2.4
Jagger	A	A	C	T	2.3	2.2
ArinaLrFor	A	A	C	T	a	a
Resistant	**A**	**A**	**C**	**T**	0.8	1.9

[a]grey highlight means the seeds were not available.

3.4.1.2 Linkage mapping confirms Wsm2 conferring WSMV resistance in 'Snowmass'

To validate the association between *Wsm2* and WSMV resistance, linkage mapping was performed in a biparental doubled-haploid mapping population (n = 116) derived from 'Snowmass' and 'Antero'. The parental cultivar 'Snowmass' is resistant to WSMV, with an average phenotype score of 0.8, whereas 'Antero' is susceptible, with a mean score of 3.3 (Table S3.1). Four significant QTL for WSMV resistance were identified by linkage mapping (LOD > 3, $P < 0.001$) on chromosomes 3B, 3D, 5B and 7B (Figure 3.2). The strongest association was identified on chromosome 3B, where 60 significant markers (LOD > 3) were mapped, including 45 within the region between 11.9 Mbp to 28.5 Mbp, co-located with the previously defined *Wsm2* region (16.5 Mbp to 18.9 Mbp) (Table S3.7). In addition, two other significant markers (LOD > 3) were mapped to chromosome 3D at physical positions 4,397,505 bp and 5,446,355 bp. Sequence alignment showed that this region was not homoeologous to the *Wsm2* locus on chromosome 3B (Figure

S3.3). Five other significant markers (LOD >3) were mapped to chromosome 7B and another significant marker (LOD >3) was mapped to chromosome 5B, respectively (Table S3.7). The results confirmed that 'Snowmass' contains the *Wsm2* variant and the association between the *Wsm2* locus and WSMV resistance at temperatures below 18 °C, making 'Snowmass' suitable plant material to characterize the genetic variants underlying the *Wsm2* locus.

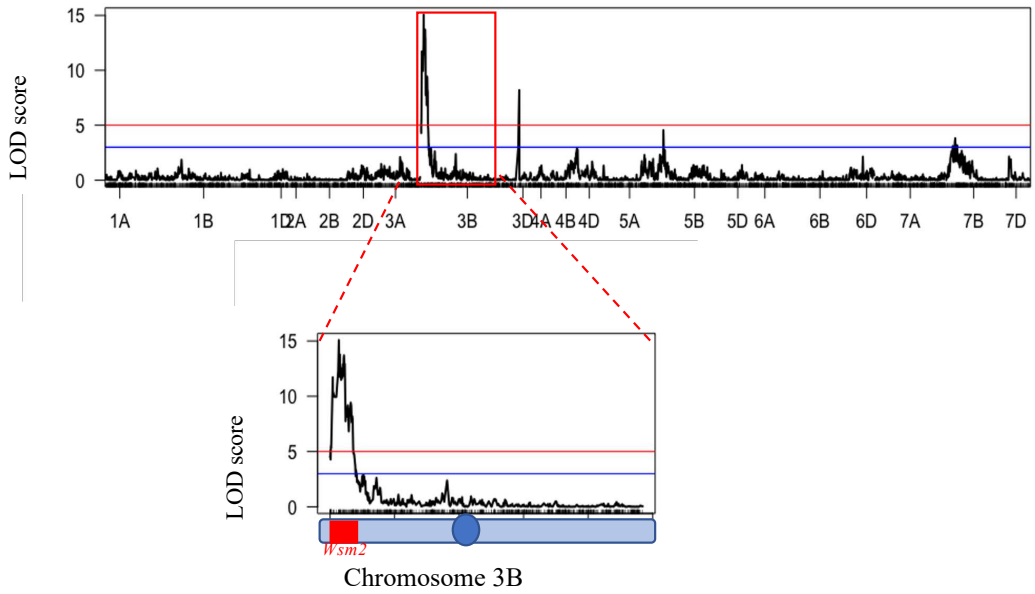


Figure 3.2. QTL mapping of doubled haploid population (n = 116) from 'Snowmass' and 'Antero'.

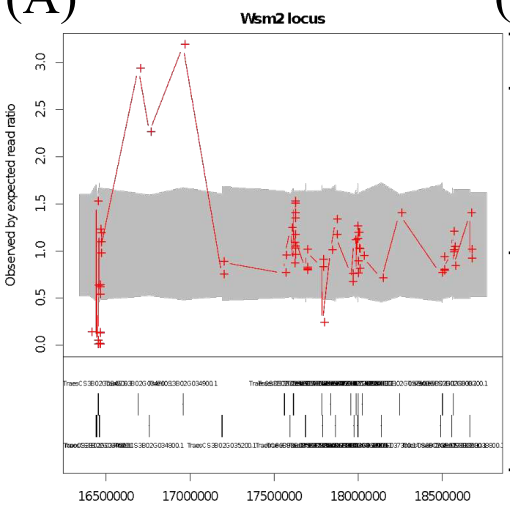
3.4.1.3 Exome capture reads revealed genetic and copy number variation underlying Wsm2

As 'Snowmass' was confirmed to carry the rare *Wsm2* allele, exome capture reads from 'Snowmass' were used to characterize natural genetic variants underlying *Wsm2*. The exome reads from 'Snowmass' were mapped to the IWGSC RefSeqv1.0 genome and their effects on protein coding genes were predicted. Within the *Wsm2* interval (16.5 Mbp to 18.9 Mbp), 851 polymorphisms were identified between 'Snowmass' exome sequences and IWGSC RefSeqv1.0 reference, leading to 3,380 genetic effects on all spliced transcript sequences (Table S3.8.1). Most

variant effects were either upstream (1,115, 33%) or downstream (1,371, 40.6%) of genes (Table S3.8.1). Six variants were predicted to have high impact on the protein coding gene, leading to either premature introduction of a stop codon within the coding sequence or frameshift variants (Table S3.8.2). Of the six high impact variants in 'Snowmass', four were also present in the WSMV susceptible parent 'Antero' (Table S3.8.2). The two high impact genetic variants unique to 'Snowmass' were insertion and deletion variants (Indels), one causing a frameshift in *TraesCS3B02G042400LC* (Patain), and another leading to a frameshift and splicing variant in *TraesCS3B02G038300* (Bowman-birk trypsin inhibitor) (Table S3.8.2).

To search for structural variation within the *Wsm2* interval in 'Snowmass', exome capture read depth from 'Snowmass' was compared to the mean depth from the exomes of six other wheat cultivars ('Antero', 'Brawl', 'Byrd', 'Hatcher', 'C0940610', and 'Platte'). None of the six cultivars exhibit resistance to WSMV except 'Hatcher' that is tolerant to WSMV (Albrecht et al., 2020). Of the 94 genes within *Wsm2*, a cluster of four adjacent UDP-glycosyltransferase genes *(TraesCS3B02G034700, TraesCS3B02G034800, TraesCS3B02G034900,* and *TraesCS3B02G035000*) also underwent copy number duplications and have two to three additional copies in 'Snowmass' compared to the mean coverage in the reference genome set. In contrast, the WD repeat-containing protein 1 (*TraesCS3B02G034400*), ubiquitin-conjugating enzyme E2 (*TraesCS3B02G034500*), and a trypsin inhibitor (*TraesCS3B02G036200*) are predicted to have fewer copies in 'Snowmass' (Figure 3.3). Collectively, the analysis of wheat pangenomes suggested that *Wsm2* is likely to be absent from wheat cultivars with genomic resources, but it's presence was confirmed in 'Snowmass'. The exome reads from 'Snowmass' showed genetic variations, such as SNP, Indels and CNVs, underlying *Wsm2* locus.

(A)



(B)

	Gene ID	Protein Annotations
Deleted Genes	TraesCS3B02G034400	WD repeat-containing protein 1
	TraesCS3B02G034500	Ubiquitin-conjugating enzyme E2
	TraesCS3B02G036200	Trypsin inhibitor
Duplicated Genes	TraesCS3B02G034700	Glycosyltransferase
	TraesCS3B02G034800	Glycosyltransferase
	TraesCS3B02G034900	Glycosyltransferase
	TraesCS3B02G035000	Glycosyltransferase

Figure 3. 3 Exome capture analysis of genomic variations underlying *Wsm2* locus in 'Snowmass'. **A)** CNV analysis for the *Wsm2* locus in 'Snowmass' compared to a wheat exome reference set composed of 'Antero', 'Brawl', 'Byrd', 'Hatcher', 'C0940610', and 'Platte' using ExomeDepth R package **B)** List of genes showing CNV, genes highlighted in red are predicted to have fewer copies in 'Snowmass', whereas genes highlighted in green are predicted to have duplicated copies in 'Snowmass'.

3.4.2 Transcriptomics to characterize plant response to WSMV infection

3.4.2.1 WSMV accumulation following inoculation in Wsm2+ and Wsm2- wheat

To compare the accumulation of WSMV in resistant and susceptible wheat, RNA was extracted from whole leaf tissues from 'Snowmass' (*Wsm2*+, WSMV resistant) and 'Antero' (*Wsm2*-, WSMV susceptible) at four time points after WSMV inoculation (0, 5, 10, and 15 dpi, Figure 3.4A). Following inoculation, WSMV accumulated in both genotypes throughout the time course, but at a much lower rate in *Wsm2*+ compared to *Wsm2*- (Figure 3.4A). There was no significant difference between genotypes in viral particles at either 0 and 5 dpi ($P > 0.05$), but at both 10 dpi (4.4-fold, $P < 0.05$) and 15 dpi (4.7-fold, $P < 0.05$) *Wsm2*+ contained significantly lower levels of WSMV compared to *Wsm2*- (Figure 3.4A). This result was consistent with visual symptoms, where *Wsm2*- individual plants showed characteristic streaked and mosaic patterns on

their leaves beginning at 10 dpi compared to *Wsm2*+ leaves that remained asymptomatic throughout the time course (Figure 3.4B).

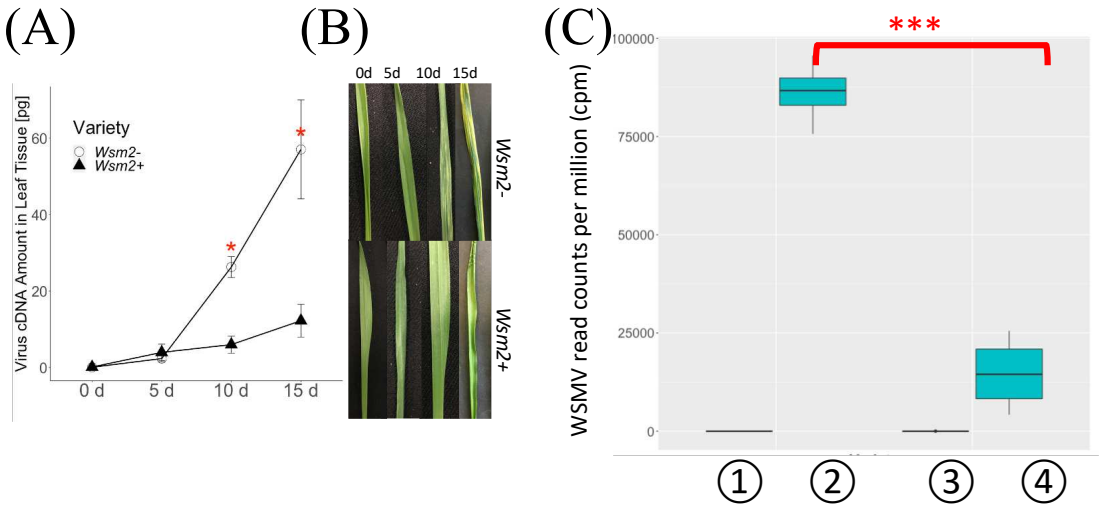


Figure 3.4 Characterization of the response of *Wsm2*+ and *Wsm2*- wheat to WSMV infection. **A)** WSMV quantification in 'Snowmass' (*Wsm2*+) and 'Antero' (*Wsm2*-) before (0), 5-, 10- and 15- day post inoculation (dpi). **B)** Phenotype of leaves in 'Snowmass' and 'Antero' 0, 5, 10, and 15- dpi. **C)** Quantification of WSMV reads in RNA-seq samples under four conditions: ①*Wsm2*- (C), ②*Wsm2*- (T), ③*Wsm2*+ (C) and ④*Wsm2*+ (T).

3.4.2.2 Summary statistics of the RNA-seq experiment

Based on the time course results, 10 dpi was selected as the time point to characterize the early transcriptomic response of wheat plants to WSMV infection. In this RNA-seq study, 16 samples were collected from two genotypes (*Wsm2*+, *Wsm2*-) and two treatments (WSMV treated, mock treated). After adaptor trimming and removal of low-quality reads, an average of 26.7 million reads were retained (Table S3.9). After alignment to the joint wheat-WSMV reference, WSMV reads were detected in all samples and calculated as WSMV read counts per million RNA seq reads (cpm) (Table S3.10). WSMV levels were much higher in WSMV-treated (T) samples (mean 50,418 cpm) compared to mock treated (C) samples (mean 5 cpm, Figure 3.4C). Comparing virus inoculated samples in both genotypes, the average amount of WSMV in *Wsm2*- (T) samples

is 5.86-fold higher compared to *Wsm2*+ (T) samples ($P < 0.001$, Figure 3.4C) from RNA-seq reads, consistent with the 4.4-fold difference in WSMV accumulation between genotypes at 10 dpi as measured by qRT-PCR (Figure 3.4A).

Overall, an average 97.2% overall mapping rate was acquired across 16 samples to the joint reference of IWGSC RefSeq v1.0 and WSMV genomes. The average unique mapping rate for *Wsm2*- wheat samples was 84.5 ± 3.3%, whereas for *Wsm2*+ samples the rate was 81.0 ± 1.3% (Table S3.9). In the sample principal component analysis (PCA), PC1 explained 45% of variance in the overall transcriptome between samples and most samples were separated according to treatment (Figure S3.4). There were three ambiguous samples, of which two belong to *Wsm2*+ (T) (R2 and R4) and one belongs to *Wsm2*- (C) (R4) (Figure S3.4). The counts of WSMV reads in the two ambiguous *Wsm2*+ (T) samples were 4,227 cpm and 9,623 cpm compared to an average of 22,440 cpm in other *Wsm2*+ (T) samples (Table S3.10), indicating that transcriptome variation in these samples may be due to variation in WSMV inoculation and infection. Samples were not grouped according to their genotype, indicating that the variation in overall transcript levels between genotypes within treatment groups were comparatively smaller compared to the treatment effect.

3.4.2.3 Host overall transcriptomic changes comparing WSMV vs. mock-treated condition

To characterize host transcriptomic responses to WSMV infection, differentially expressed genes between mock and WSMV-treated susceptible materials (*Wsm2*-) were analyzed ($Wsm2\ -_C^T$) (Table S3.11). In total, 8,975 genes were identified as differentially expressed genes (DEGs) in $Wsm2\ -_C^T$, of which 5,031 genes were up-regulated and 3,944 genes were down-regulated after WSMV infection (Figure 3.5A). This comprised 74.2% of the total 121,016 expressed genes,

defined as raw read counts ≥ 1 in at least one of the samples, indicating that host plants undergo major transcriptional reprogramming in response to WSMV infection.

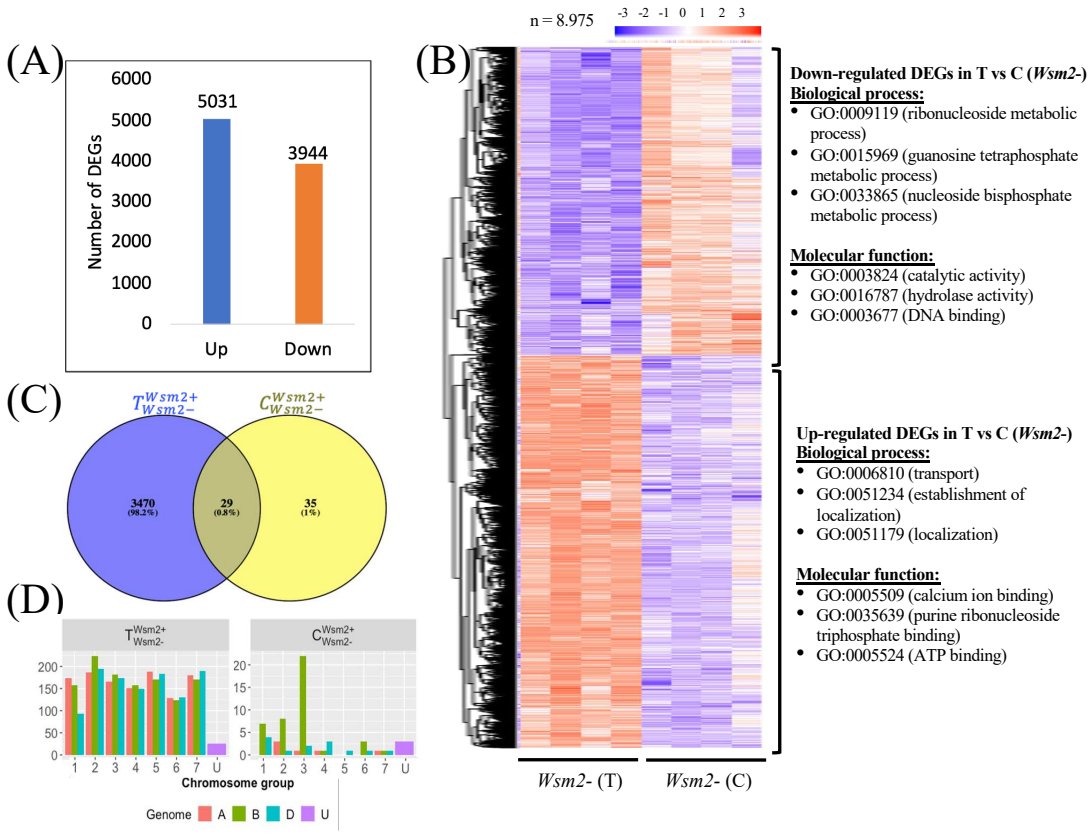


Figure 3.5 Overview of differentially expressed genes between treatment and between genotype (T^{Wsm2+}_{Wsm2-} and C^{Wsm2+}_{Wsm2-}) using IWGSC RefSeq v1.0 as a mapping reference. **a)** Number of up- and down-regulated DEGs (*P*adj < 0.01) between treatment (WSMV treated *vs.* Mock treated samples. **b)** Heatmap of 8,975 DEGs from comparison of WSMV-treated *vs.* mock treated *Wsm2*- samples. The expression values are normalized by setting the mean of every row to 0 and the standard deviation of every row to 1. Hierarchical clustering separated these into DEGs that are either upregulated (n=5,031) or downregulated (n=3,944) in WSMV treated conditions. The top three enriched GO terms (biological process and molecular function) for each row cluster are shown on the right. **c)** Venn diagram of total DEGs (*P*adj < 0.01) between genotypes. **d)** Number of DEGs comparing *Wsm2*+ vs. *Wsm2*- under WSMV treated (T^{Wsm2+}_{Wsm2-}) or mock treated (C^{Wsm2+}_{Wsm2-})conditions, located on each chromosome, genomes are color coded.

Genes that were up-regulated in response to WSMV infection were significantly enriched (*P* < 0.01) for biological process GO terms related to 'transport' (GO:0006810) and 'localization' (GO:0051179), and for the molecular function GO terms 'binding to calcium ion' (GO:0005509), 'protein binding' (GO:0005515), and 'ATP binding' (GO:0005524) (Figure 3.5B, Table S3.12),

indicating these processes are activated in host plants in response to viral infection. By contrast, genes down-regulated in response to WSMV treatment were most significantly enriched for biological process GO terms relating to 'metabolic processes' (GO:0015969, GO:0009119, GO:0033865) and the molecular function GO terms 'catalytic or hydrolase activity' (GO:0003825, and GO:0016787) (Figure 3.5B, Table S3.12), indicating host plants suppress metabolic activity in response to infection.

3.4.2.4 Comparing transcriptional differences between genotypes under mock and WSMV-treated condition

To characterize transcriptional changes between genotypes, DEGs were analyzed under mock inoculation (C_{Wsm2-}^{Wsm2+}) and virus inoculation (T_{Wsm2-}^{Wsm2+}) conditions. Sixty-four genes were differentially expressed between genotypes in the C_{Wsm2-}^{Wsm2+} comparison (Figure 3.5C), of which 28 genes were up-regulated, and 36 were down-regulated in the *Wsm2*+ genotype (Table S3.11). Of the 64 DEGs, 22 (34.4%) are located on chromosome 3B (Figure 3.5D), with 14 close to *Wsm2* markers (from 13.7 Mb to 30.5 Mb) and five of the 94 genes defined in the region of the IWGSC RefSeq v1.0 genome (Table S3.13).

In comparison, 3,499 genes were differentially expressed between genotypes under WSMV-treated condition (T_{Wsm2-}^{Wsm2+}), of which 1,920 were up-regulated and 1,579 were down-regulated in *Wsm2*+ genotypes. Among the 3,499 DEGs in T_{Wsm2-}^{Wsm2+}, 2,059 (58.8%) of them have less than two-fold-change values in one genotype versus the other (Table S3.11). Twenty-nine genes were differentially expressed between genotypes in both mock and WSMV-treated conditions, while 3,470 genes were differentially expressed only after virus treatment (Figure 3.5C). These results are consistent with the PCA plot, indicating that WSMV infection induces a major shift in the transcriptome profile of the host plant. These 3,470 DEG are significantly enriched for GO terms

relating to different metabolic process (GO:0046128, GO:0072521, GO:0033865) and catalytic activity (GO:0003824, GO:0016757) (Table S3.12), indicating the *Wsm2*-mediated resistance likely causes major transcriptional changes on metabolic process or catalytic activity at 10 dpi. However, examination of enriched GO terms for these DEGs did not find any 'defense response', 'hormone regulation', or 'signaling transduction' related terms (Table S3.12), suggesting resistant responses may not be significant at this time point.

3.4.3 Analysis of transcriptomes to identify causative genes underlying *Wsm2*

3.4.3.1 De novo assembly of unmapped reads revealed three unique transcripts in Wsm2+

To identify genetic variants or causative genes underlying *Wsm2* that are specific to *Wsm2+* materials and absent in the wheat reference genome IWGSC RefSeq v1.0 (*Wsm2-*, WSMV susceptible), a *de novo* assembly of RNA-seq reads that did not map to this reference was performed. A total of 23,066,200 unmapped reads (5.4% of all reads, Table S3.9) were combined from all 16 samples and assembled into 175,897 transcripts. Using CD-HIT clustering, these were collapsed to 161,210 non-redundant transcripts for subsequent analysis.

The unmapped RNA-seq reads from all samples were mapped back to the *de novo* assembled transcriptomes to identify differentially expressed transcripts in *Wsm2+* (T) compared to both *Wsm2-* (T) and *Wsm2+* (C) conditions. From the two pairwise comparisons, 84 transcripts were differentially expressed between *Wsm2+* (T) and *Wsm2-* (T), in addition to 102 transcripts that were differentially expressed between *Wsm2+* (T) and *Wsm2+* (C) (Table S3.14). Fourteen transcripts identified from unmapped reads were differentially expressed in both comparisons, 11 of which were annotated as WSMV polyprotein gene, indicating those reads likely represent WSMV transcripts from the inoculum that have replicated in plant leaf tissues (Table S3.15). For the remaining three transcripts that were differentially expressed in both comparisons, one was

annotated as leaf rust 10 disease resistance locus receptor-like protein kinase (XM_037561459.1) and was significantly up-regulated in *Wsm2*+ (T) condition (TPM = 5.5, $P < 0.001$), but not expressed in either *Wsm2*- (T) or *Wsm2*+ (C) conditions. The other two were annotated as lectin-like receptor kinase gene (MT027257.1), and cytochrome P450 (XM_037560405.1), and were expressed at low levels (TPM < 1.5) in *Wsm2*+ (C) condition and after WSMV infection they were significantly downregulated (5.7-fold change, $P < 0.001$) in T_{Wsm2-}^{Wsm2+} (Table S3.15), indicating they likely promote plant susceptibility to WSMV infection.

To confirm these three transcripts were truly absent in the wheat reference genome, their coding sequences were used as queries in BLASTn searches against a database of high and low confidence transcripts from the IWGSC RefSeq v1.0 assembly. The sequence similarity between these transcripts and the top BLAST hit in IWGSC RefSeq v1.0 was less than 96% (Table S3.16), validating this approach using *de novo* assembly of unmapped reads to detect the non-wheat reference genome annotated transcripts.

To study if these transcripts were unique in *Wsm2*+ material, their sequences were also used as queries in BLASTn searches against the coding sequence (CDS) of ten other wheat cultivars. The cytochrome P450 transcript was present on the distal end of the long arm of chromosome 3B in all ten wheat cultivars near 840 Mbp, far from the *Wsm2* locus that is located on the distal end of short arm 3B from 16.5 Mbp to 18.9 Mbp, suggesting it is unlikely to be orthologous to the *Wsm2* region (Table S3.16). The leaf rust 10 disease resistance locus transcript was present in six wheat cultivars, and in each case located on chromosome 3B approximately 20 Mbp downstream of the KASP1 marker for the *Wsm2* locus (Table S3.16). The lectin-like receptor kinase transcript was absent in all wheat cultivars, indicating that this likely represents a rare transcript in *Wsm2*+ materials (Table S3.16).

3.4.3.2 Examination of candidate genes within Wsm2 found five possible causative genes underlying Wsm2

To maximize the possibility to detect all possible *Wsm2* causative genes, the candidate genes within *Wsm2* interval based on wheat reference genome were studied as well. Of the 94 annotated candidate genes underlying *Wsm2* interval from IWGSC RefSeq v1.0, five were differentially expressed between *Wsm2*+and *Wsm2*- at 10 dpi in the RNA-seq analysis and considered as possible causative genes underlying *Wsm2* (Figure 3.6A). Their expression changes were validated using qRT-PCR (Figure 3.6B), demonstrating the reliability of RNA-seq to quantify transcript levels. Of the five candidates, two were more highly expressed in *Wsm2*- and encode WD repeat-containing protein 1 (*TraesCS3B02G034400*) and Ubiquitin-conjugating enzyme E2 (*TraesCS3B02G034500*) (Table S3.17). Consistently, the exome CNV analysis for these two candidates (*TraesCS3B02G034400, TraesCS3B02G034500*) revealed they have reduced copy number in 'Snowmass' (Figure 3.3), demonstrating a link between decreased transcript levels with a reduction in genome copy number for these two candidates in *Wsm2*+.

The other three candidate genes were more highly expressed in *Wsm2*+ genotypes under both WSMV-treated and mock treated conditions, and encode SUF system FeS (*TraesCS3B02G035600*), Disease resistance protein RPM1 (*TraesCS3B02G035800*), and Chaperone protein DnaK (*TraesCS3B02G035900*). However, none of these up-regulated candidates in *Wsm2*+ were found with increased copy numbers in 'Snowmass' from the exome capture analysis (Figure 3.3), indicating their higher transcript levels in *Wsm2*+ is likely due to genetic variation in the promoter or other regulatory region.

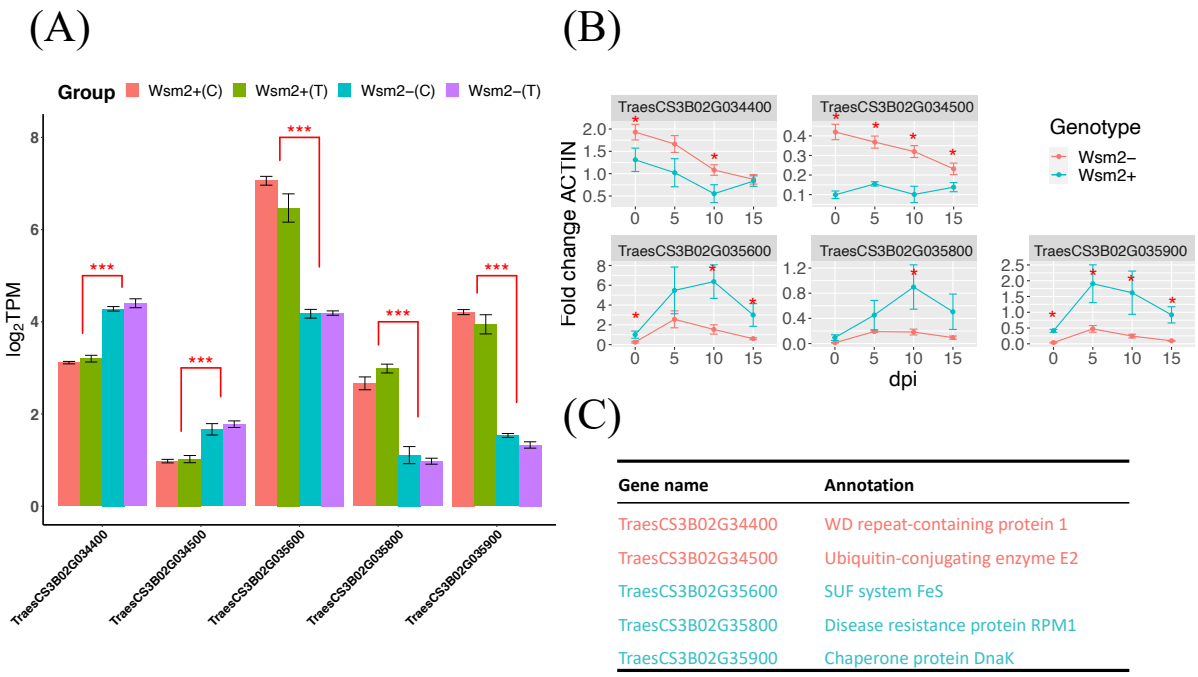


Figure 3.6. Transcript levels of five DEGs within *Wsm2* region. **(A)** log_2TPM values of these five DEGs within *Wsm2* interval from RNA-seq study. They were color-coded by four groups, *Wsm2*+ under WSMV treated (*Wsm2*+ T) and mock treated condition (*Wsm2*+ C) as well as *Wsm2*- under WSMV treated (*Wsm2*- T) and mock treated condition (*Wsm2*- C). **(B)** Expression in fold change to *ACTIN* reference for these five DEGs in a time course experiment with four time points (0, 5, 10, and 15-dpi). Samples were from leaves in *Wsm2*+ and *Wsm2*- after WSMV infection and RNA were extracted and run with qRT-PCR. * = $P < 0.05$. **(C)** Annotation for the five DEGs

Of the four UDP-glycosyltransferase genes predicted to exhibit increased copy number in 'Snowmass', three showed no significant differences in expression between genotypes and one gene (*TraesCS3B02G034800*) had unexpectedly lower expression in *Wsm2*+ samples in both WSMV treated and mock conditions ($P < 0.05$, Table S3.17). This result suggests that the increased copies likely do not play a role in plant response to viral or mock infection conditions.

Previously, the Bowman-Birk inhibitor (BBI) gene family, with potential roles in biotic stress resistance, was found to have undergone extensive copy number and domain number duplication within the *Wsm2* region (Xie et al., 2021). However, none of the fourteen BBIs located on chromosome 3B within the *Wsm2* region were differentially expressed between genotypes (Table S3.17). Twelve of these genes have nearly 0 TPM across all samples, suggesting they are unlikely

to be induced or suppressed by WSMV infection at 10 dpi. Collectively, five of the 94 candidate genes within *Wsm2* region, which were differentially expressed between genotypes, were considered as potential causative genes underlying *Wsm2*. One of the candidates annotated as the Disease resistance protein RPM1 (*TraesCS3B02G035800*) encodes an LRR domain protein was considered as a top candidate and was subjected to further functional validation.

3.4.3.3 Generation of CRISPR/Cas9 edited mutant for disease resistance protein RPM1

To characterize the function of wheat *RPM1* (*TraesCS3B02G035800*) in WSMV resistance, the CRISPR/Cas9 genome editing tool was applied to generate gene knockout mutant for functional validation. The coding sequence of *TraesCS3B02G035800* was used as a query in a BLASTn search against the IWGSC RefSeq v1.0 genome, revealing one homoeologous gene on chromosome 3D (*TraesCS3D02G032900*). There was no homoeologue on chromosome 3A, either in 'Chinese Spring', or ten other wheat cultivars that were analyzed (Table S3.18), suggesting this homoeologue has been lost in modern wheat during evolution.

Two CRISPR/Cas9 constructs with distinct sgRNAs designed to target both B and D homoeologues were assembled and transformed into 'Snowmass' and 'Snowmass 2.0' embryos. Snowmass 2.0 (CO07W22-F5/Snowmass//Brawl CL Plus) has approximately one quarter of the genetic material from 'Snowmass', maintaining the *Wsm2* allele and demonstrating strong WSMV resistance. A total of 24 transgenic plants were generated, twelve in 'Snowmass' background and another twelve in 'Snowmass 2.0' background (Table S3.19). The genotyping assay confirmed none of the transgenic plants in 'Snowmass' background has CRISPR/Cas9 edits, whereas three transgenic plants (#13, #16, and #23) in 'Snowmass 2.0' background have edits for both B and D homoeologues (Figure S3.1). The transgenic plant #13 has three SNPs mutation (C110G+T114G+T117G) on B copy and one SNP insertion (111insC) on D copy, plant #16 has

one SNP insertion and two SNPs mutations (113insG+T117A+G122A) on B copy and one SNP deletion on D copy (108delC), and plant #23 has one SNP deletion for both B and D gene copies (474delT) (Table S3.19). All mutations were in a heterozygous state. These T_0 plant materials will be selfed to obtain T_1 plants carrying homozygous mutations in both B and D gene copies to characterize the function of *RPM1* using WSMV inoculation assays.

3.5 Discussion

3.5.1 *Wsm2* is highly dynamic within wheat species and likely absent from modern wheat cultivars

The *Wsm2* locus has been mapped to the telomeric region of chromosome arm 3BS (Assanga et al., 2017; Lu et al., 2012; Tan et al., 2017) within a 2.4 Mbp interval (16.5 - 18.9 Mbp) based on the wheat reference genome assembly IWGSC RefSeq v1.0 (Figure 3.1). Analyses in the current study demonstrated that the *Wsm2* region is diverse among different common wheat cultivars, with large structural variations underlying this locus within species (Figure 3.1). The telomeric regions of the chromosomes are associated with high rates of recombination during gene evolution, which results in frequent duplication and divergence events (See et al., 2006; Saintenac et al., 2009). Such high frequency in crossover and introgression at the end of chromosome 3BS are possible contributing factors for the large deletions/insertions and inversions for *Wsm2* in different wheat cultivars.

Moreover, this study showed that although *Wsm2* appears to originate in common wheat, it is likely absent from all seventeen wheat cultivars with sequenced genomes, including the wheat reference genome 'Chinese Spring' (Table 3.1). This result is in agreement with Tan et al., 2017 that showed 'Chinese Spring' is susceptible to WSMV and with Zhang and Hua, 2018 who showed that *Wsm2* is absent from wild *Brachypodium* accessions. Despite the diverse haplotypes in these

seventeen wheat cultivars, they all exhibited WSMV susceptibility (Table 3.1). The lack of association of the markers with WSMV resistance phenotypes indicated that current markers may not be predictive for *Wsm2*. Therefore, sequencing information from *Wsm2* locus carrying wheat material would be useful to understand the natural variation underlying the *Wsm2* locus and to identify perfect markers benefitting breeders.

3.5.2 Applying genomic and transcriptomic resources to characterize natural variation underlying *Wsm2*

The wheat pangenome is a powerful resource to exploit natural variations and characterize the genetic variants associated with important agronomic traits and resistance to stresses (Walkowiak et al., 2020). The identification of haplotypes associated with distinct phenotypes from different wheat cultivars, and the corresponding genomic sequences can be extracted from such sequenced cultivars and for subsequent comparative genomic analysis to identify unique variants underlying genetic variants of interest (Brinton et al., 2020). However, such an approach is limited if the genetic variants of interest are not present in those sequenced wheat germplasms. This will become less of a problem in future as more and more cultivars are sequenced, and high-quality sequence information is generated from a cultivar of interest.

To study the genetic variants from a cultivar of interest, targeted sequencing such as chromosome arm sequencing has been successfully applied to clone the broad-spectrum *Lr22a* leaf-rust resistance gene (Thind et al., 2017). However, chromosome sequencing remains costly. Instead, exome capture reads, genotyping-by-sequencing (GBS), or transcriptome data generated from a cultivar of interest are important resources to reduce the cost and complexity of whole genomes but still able to study critical genetic variants.

Our study validated the presence of *Wsm2* in 'Snowmass' using GBS markers and linkage mapping in a doubled haploid population (Figure 3.2). We further analyzed exome reads in 'Snowmass' and compared genetic and structural variants underlying *Wsm2* with wheat reference genome IWGSC RefSeq v1.0 (Figure 3.3 and Table S3.8). Significant copy number duplications for four tandem duplicated UDP-glycotransferases genes were identified in 'Snowmass' (Figure 3.3). The UDP-glycosyltransferase gene plays diverse roles in plant immunity against various types of pathogens, for example, it functions as a negative regulator of the necrotrophic fungus *Botrytis cinerea* (Castillo et al., 2019), whereas it promotes resistance to the hemi-biotrophic bacterial pathogen *Pseudomonas syringae* pv *tomato* carrying the *AvrRpm1* gene (Langlois-Meurinne et al., 2005).

The transcriptome reads from *Wsm2*+ wheat materials were mapped to IWGSC RefSeq v1.0 and the unmapped reads were subjected to *de novo* transcriptome assembly, from which three unique transcripts specific to *Wsm2*+ were identified (leaf rust 10 disease resistance locus protein kinase, lectin-like receptor kinase, and cytochromes P450, Table S3.15). Cytochromes P450 proteins are known to function in phytoalexin biosynthesis, hormone metabolism regulation, and the biosynthesis of secondary metabolites and other defensive signaling molecules which regulate plant immunity against various pathogen types (Xu et al., 2015). The lectin-like receptor kinase gene (*LecRLK*) is a class of RLK that has many copies present on plant genomes and contains a lectin/lectin-like ectodomain which can bind to carbohydrate (Sun et al., 2020). *LecRLKs* are involved in plant basal defense against both biotrophic and necrotrophic pathogens through carbohydrate signal perception which triggers the PTI response (Sun et al., 2020). However, whether *LecRLKs* are also involved in ETI, or if they play a role in plant response to viral infection, remains unknown. The gene annotated as leaf rust 10 disease-resistance locus receptor-like protein

kinase-like (*LRK10*) was first identified from wheat providing resistance to fungal pathogen *Puccinia triticina* (Feuillet et al., 1997, 1998). The *LRK10* gene was later characterized as an NLR-class of *R* gene in wheat with a strong diversifying selected N-terminal CC domain, suggesting a complex molecular mechanism of pathogen detection and signal transduction (Loutre et al., 2009). Although no evidence suggests this *LRK10* is involved in plant defense response to viral pathogens, it is possible that this candidate could directly or indirectly interact with viral molecules and be involved in a downstream signal transduction pathway important in immunity.

3.5.3 Transcriptomics revealed possible causative genes underlying *Wsm2*

We also examined annotated genes within *Wsm2* interval based on wheat reference genome IWGSC RefSeq v1.0 for those differentially expressed between genotypes and identified five possible candidates (Figure 3.6). The two down-regulated DEGs in *Wsm2*+ are annotated as WD repeat-containing protein 1 (*TraesCS3B02G034400*) and ubiquitin-conjugating enzyme E2 (*TraesCS3B02G034500*), which are members of large protein families common in all eukaryotes with functions associated with basic cellular mechanisms (van Nocker & Ludwig, 2003).

Two of the up-regulated DEGs in *Wsm2*+ are annotated as SUF system FeS (*TraesCS3B02G035600*) and Chaperone protein DnaK (*TraesCS3B02G035900*) (Figure 3.6). The chaperone protein DnaK (HSP70) is known to respond to both biotic and abiotic stress, by which it helps to prevent the accumulation of excessive newly synthesized proteins, and to ensure proper protein folding during their transition process (Park & Seo, 2015). Another up-regulated candidate in *Wsm2*+, annotated as *RPM1* (*TraesCS3B02G035800*), is a known plant resistance (*R*) gene and encodes a CC-NB-LRR domain protein that can recognize the avirulence factor AvrRpm1 from the bacterial pathogen *Pseudomonas syringae* pv. *maculicola* 1 and trigger plant ETI defense responses (Grant et al., 1995). We prioritized this *RPM1* gene as a strong candidate gene because

of its known role in resistance to pathogens in other plant species and used the CRISPR/Cas9 genome editing tool with GRF-GIF system to generate gene knockout mutant in ‘Snowmass 2.0’. These mutant plants will be valuable materials to characterize the function of *RPM1* in WSMV resistance.

In conclusion, we demonstrated *Wsm2* is a rare allele in modern common wheat cultivars and is likely to have unique genetic variation in ‘Snowmass’. We applied genomic and transcriptomic tools to characterize variants in ‘Snowmass’ and other *Wsm2*+ lines and searched for evidence to identify the possible causative genes underlying this *Wsm2* locus. We selected one of the top candidates and generated gene knockout mutants using CRISPR/Cas9 technology.

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., & Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research, 46(W1), W537–W544. https://doi.org/10.1093/nar/gky379

Albrecht, T., White, S., Layton, M., Stenglein, M., Haley, S., & Nachappa, P. (2020). Ecology and epidemiology of wheat curl mite and mite-transmissible viruses in Colorado and insights into the wheat virome. BioRxiv, 2020.08.10.244806. https://doi.org/10.1101/2020.08.10.244806

Alexa A, R. J. (2021). TopGO: Enrichment Analysis for Gene Ontology.

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C. J., Choulet, F., Distelfeld, A., Poland, J., Ronen, G., Barad, O., Baruch, K., Keeble-Gagnère, G., Mascher, M., Ben-Zvi, G., Josselin, A. A., Himmelbach, A., Balfourier, F., … Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 361(6403), 7191–7191. https://doi.org/10.1126/science.aar7191

Arends, D., Prins, P., Jansen, R. C., & Broman, K. W. (2010). R/qtl: High-throughput multiple QTL mapping. Bioinformatics (Oxford, England), 26(23), 2990–2992. https://doi.org/10.1093/bioinformatics/btq565

Assanga, S., Zhang, G., Tan, C. T., Rudd, J. C., Ibrahim, A., Xue, Q., Chao, S., Fuentealba, M. P., & Liu, S. (2017). Saturated genetic mapping of wheat streak mosaic virus resistance gene *Wsm2* in wheat. Crop Science, 57(1), 332–339. https://doi.org/10.2135/cropsci2016.04.0233

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brinton, J., Ramirez-Gonzalez, R. H., Simmonds, J., Wingen, L., Orford, S., Griffiths, S., Haberer, G., Spannagl, M., Walkowiak, S., Pozniak, C., & Uauy, C. (2020). A haplotype-led approach to increase the precision of wheat breeding. Communications Biology, 3(1). https://doi.org/10.1038/s42003-020-01413-2

Castillo, N., Pastor, V., Chávez, Á., Arró, M., Boronat, A., Flors, V., Ferrer, A., & Altabella, T. (2019). Inactivation of UDP-glucose sterol glucosyltransferases enhances *Arabidopsis* resistance to *Botrytis cinerea*. Frontiers in Plant Science, 10, 1162–1162. https://doi.org/10.3389/fpls.2019.01162

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chisholm, S. T., Parra, M. A., Anderberg, R. J., & Carrington, J. C. (2001). *Arabidopsis RTM1* and *RTM2* genes function in phloem to restrict long-distance movement of tobacco etch virus. Plant Physiology, 127(4), 1667–1675. https://doi.org/10.1104/pp.010479

Chuang, W. P., Rojas, L. M. A., Khalaf, L. K., Zhang, G., Fritz, A. K., Whitfield, A. E., & Smith, C. M. (2017). Wheat genotypes with combined resistance to wheat curl mite, wheat streak mosaic virus, wheat mosaic virus, and *Triticum* mosaic virus. Journal of Economic Entomology, 110(2), 711–718. https://doi.org/10.1093/jee/tow255

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide

polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly, 6(2), 80–92. https://doi.org/10.4161/fly.19695

Cosson, P., Schurdi-Levraud, V., Le, Q. H., Sicard, O., Caballero, M., Roux, F., Le Gall, O., Candresse, T., & Revers, F. (2012). The *RTM* resistance to potyviruses in *Arabidopsis thaliana*: natural variation of the *RTM* genes and evidence for the implication of additional genes. PLOS ONE, 7(6), e39169.

Cosson, P., Sofer, L., Le, Q. H., Léger, V., Schurdi-Levraud, V., Whitham, S. A., Yamamoto, M. L., Gopalan, S., le Gall, O., Candresse, T., Carrington, J. C., & Revers, F. (2010). *RTM3*, which controls long-distance movement of potyviruses, is a member of a new plant gene family encoding a meprin and TRAF homology domain-containing protein. Plant Physiology, 154(1), 222–232. https://doi.org/10.1104/pp.110.155754

Cram, D., Kulkarni, M., Buchwaldt, M., Rajagopalan, N., Bhowmik, P., Rozwadowski, K., Parkin, I. A. P., Sharpe, A. G., & Kagale, S. (2019). WheatCRISPR: A web-based guide RNA design tool for CRISPR/Cas9-mediated genome editing in wheat. BMC Plant Biology, 19(1), 474–474. https://doi.org/10.1186/s12870-019-2097-z

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2). https://doi.org/10.1093/gigascience/giab008

Danilova, T. V., Zhang, G., Liu, W., Friebe, B., & Gill, B. S. (2017). Homoeologous recombination-based transfer and molecular cytogenetic mapping of a wheat streak mosaic virus and *Triticum* mosaic virus resistance gene *Wsm3* from *Thinopyrum intermedium* to wheat. Theoretical and Applied Genetics, 130(3), 549–556. https://doi.org/10.1007/s00122-016-2834-8

de Ronde, D., Butterbach, P., & Kormelink, R. (2014). Dominant resistance against plant viruses. Frontiers in Plant Science, 5(JUN), 1–17. https://doi.org/10.3389/fpls.2014.00307

Debernardi, J. M., Tricoli, D. M., Ercoli, M. F., Hayta, S., Ronald, P., Palatnik, J. F., Dubcovsky, J. (2020). A GRF–GIF chimeric protein improves the regeneration efficiency of transgenic plants. Nature Biotechnology, 38(11), 1274–1279. https://doi.org/10.1038/s41587-020-0703-0

Dhakal, S., Tan, C. T., Anderson, V., Yu, H., Fuentealba, M. P., Rudd, J. C., Haley, S. D., Xue, Q., Ibrahim, A. M. H., Garza, L., Devkota, R. N., & Liu, S. (2018). Mapping and KASP marker development for wheat curl mite resistance in "TAM 112" wheat using linkage and association analysis. Molecular Breeding, 38(10), 119–119. https://doi.org/10.1007/s11032-018-0879-x

Divis, L. A., Graybosch, R. A., Peterson, C. J., Baenziger, P. S., Hein, G. L., Beecher, B. B., & Martin, T. J. (2006). Agronomic and quality effects in winter wheat of a gene conditioning resistance to wheat streak mosaic virus. Euphytica, 152(1), 41–49. https://doi.org/10.1007/s10681-006-9174-8

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. Current Protocols in Bioinformatics, 51(1), 11.14.1-11.14.19. https://doi.org/10.1002/0471250953.bi1114s51

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE, 6(5), e19379–e19379.

FAOSTAT. (2020). Food and Agriculture Organization of the United Nations Database, FAOSTAT statistics, Crops-FAOSTAT statistics, Crops. https://doi.org/10.1016/B978-0-12-384947-2.00270-1

Feuillet, C., Reuzeau, C., Kjellbom, P., & Keller, B. (1998). Molecular characterization of a new type of receptor-like kinase (*wlrk*) gene family in wheat. Plant Molecular Biology, 37(6), 943–953. https://doi.org/10.1023/A:1006062016593

Feuillet, C., Schachermayr, G., & Keller, B. (1997). Molecular cloning of a new receptor-like kinase gene encoded at the *Lr10* disease resistance locus of wheat. Plant Journal, 11(1), 45–52. https://doi.org/10.1046/j.1365-313X.1997.11010045.x

Friebe, B., Qi, L. L., Wilson, D. L., Chang, Z. J., Seifers, D. L., Martin, T. J., Fritz, A. K., & Gill, B. S. (2009). Wheat-*thinopyrum intermedium* recombinants resistant to wheat streak mosaic virus and *Triticum* mosaic virus. Crop Science, 49(4), 1221–1226. https://doi.org/10.2135/cropsci2008.09.0513

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. https://doi.org/10.1038/nbt.1883

Grant, M. R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R. W., & Dangl, J. L. (1995). Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. Science, 269(5225), 843–846. https://doi.org/10.1126/science.7638602

Haber, S., Seifers, D. L., & Thomas, J. (2006). A new source of resistance to Wheat streak mosaic virus in spring wheat. 28(2), 324–324.

Hadi, B. A. R., Langham, M. A. C., Osborne, L., & Tilmon, K. J. (2011). Wheat streak mosaic virus on wheat: Biology and management. Journal of Integrated Pest Management, 2(1), J1–J5. https://doi.org/10.1603/IPM10017

Haley, S. D., Johnson, J. J., Peairs, F. B., Stromberger, J. A., Heaton, E. E., Seifert, S. A., Kottke, R. A., Rudolph, J. B., Martin, T. J., Bai, G., Chen, X., Bowden, R. L., Jin, Y., Kolmer, J. A., Seifers, D. L., Chen, M.-S., & Seabourn, B. W. (2011). Registration of 'Snowmass' wheat. Journal of Plant Registrations, 5(1), 87–90. https://doi.org/10.3198/jpr2010.03.0175crc

Haley, S. D., Martin, T. J., Quick, J. S., Seifers, D. L., Stromberger, J. A., Clayshulte, S. R., Clifford, B. L., Peairs, F. B., Rudolph, J. B., Johnson, J. J., Gill, B. S., & Friebe, B. (2002). Registration of CO960293-2 wheat germplasm resistant to Wheat streak mosaic virus and Russian wheat aphid. Crop Science, 42(4), 1381–1382. https://doi.org/10.2135/cropsci2002.1381

Harvey, T. L., Seifers, D. L., Martin, T. J., Brown-Guedira, G., & Gill, B. S. (1999). Survival of wheat curl mites on different sources of resistance in wheat. Crop Science, 39(6), 1887–1889. https://doi.org/10.2135/cropsci1999.3961887x

Hayta, S., Smedley, M. A., Demir, S. U., Blundell, R., Hinchliffe, A., Atkinson, N., & Harwood, W. A. (2019). An efficient and reproducible Agrobacterium-mediated transformation method for hexaploid wheat (*Triticum aestivum* L.). Plant Methods, 15(1), 121–121. https://doi.org/10.1186/s13007-019-0503-z

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K., Fritz, A., Hucl, P., Wiebe, K., Knox, R., Cuthbert, R., Pozniak, C., Akhunova, A., Morrell, P. L., Davies, J. P., Webb, S. R., Spangenberg, G., Hayes, B., … Akhunov, E. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nature Genetics, 51(5), 896–904. https://doi.org/10.1038/s41588-019-0382-2

Hunger, R., Sherwood, J., Evans CK, & Montana, J. (1992). Effects of planting date and inoculation date on severity of wheat streak mosaic in hard red winter wheat cultivars. Plant Disease, 76, 1056–1060.

Ishibashi, K., & Ishikawa, M. (2013). The resistance protein Tm-1 inhibits formation of a tomato mosaic virus replication protein-host membrane protein complex. Journal of Virology, 87(14), 7933–7939. https://doi.org/10.1128/jvi.00743-13

Ishibashi, K., Masuda, K., Naito, S., Meshi, T., & Ishikawa, M. (2007). An inhibitor of viral RNA replication is encoded by a plant resistance gene. Proceedings of the National Academy of Sciences of the United States of America, 104(34), 13833–13838. https://doi.org/10.1073/pnas.0703203104

Jordan, K. W., Wang, S., Lun, Y., Gardiner, L. J., MacLachlan, R., Hucl, P., Wiebe, K., Wong, D., Forrest, K. L., Sharpe, A. G., Sidebottom, C. H. D., Hall, N., Toomajian, C., Close, T., Dubcovsky, J., Akhunova, A., Talbert, L., Bansal, U. K., Bariana, H. S., … Akhunov, E. (2015). A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biology, 16(1), 48–48. https://doi.org/10.1186/s13059-015-0606-4

Koressaar, T., & Remm, M. (2007). Enhancements and modifications of primer design program Primer3. Bioinformatics, 23(10), 1289–1291. https://doi.org/10.1093/bioinformatics/btm091

Kourelis, J., & van der Hoorn, R. A. L. (2018). Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. The Plant Cell, 30(2), 285–299. https://doi.org/10.1105/tpc.17.00579

Kumssa, T. T., Rupp, J. S., Fellers, M. C., Fellers, J. P., & Zhang, G. (2019). An isolate of Wheat streak mosaic virus from foxtail overcomes *Wsm2* resistance in wheat. Plant Pathology, 68(4), 783–789. https://doi.org/10.1111/ppa.12989

Langlois-Meurinne, M., Gachon, C. M. M., & Saindrenan, P. (2005). Pathogen-responsive expression of glycosyltransferase genes UGT73B3 and UGT73B5 is necessary for resistance to *Pseudomonas syringae* pv tomato in *Arabidopsis*. Plant Physiology, 139(4), 1890–1901. https://doi.org/10.1104/pp.105.067223

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liu, S., Assanga, S. O., Dhakal, S., Gu, X., Tan, C.-T., Yang, Y., Rudd, J., Hays, D., Ibrahim, A., Xue, Q., Chao, S., Devkota, R., Shachter, C., Huggins, T., Mohammed, S., & Fuentealba, M. P. (2016). Validation of chromosomal locations of 90K array single nucleotide polymorphisms in US wheat. Crop Science, 56(1), 364–373. https://doi.org/10.2135/cropsci2015.03.0194

Liu, W., Seifers, D. L., Qi, L. L., Friebe, B., & Gill, B. S. (2011). A compensating wheat–*thinopyrum intermedium* robertsonian translocation conferring resistance to wheat streak mosaic virus and *Triticum* mosaic virus. Crop Science, 51(6), 2382–2390. https://doi.org/10.2135/cropsci2011.03.0118

Loutre, C., Wicker, T., Travella, S., Galli, P., Scofield, S., Fahima, T., Feuillet, C., & Keller, B. (2009). Two different CC-NBS-LRR genes are required for *Lr10*-mediated leaf rust

resistance in tetraploid and hexaploid wheat. Plant Journal, 60(6), 1043–1054. https://doi.org/10.1111/j.1365-313X.2009.04024.x

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550–550. https://doi.org/10.1186/s13059-014-0550-8

Lu, H., Kottke, R., Devkota, R., Amand, P. S., Bernardo, A., Bai, G., Byrne, P., Martin, T. J., Haley, S. D., & Rudd, J. (2012). Consensus mapping and identification of markers for marker-assisted selection of *Wsm2* in wheat. Crop Science, 52(2), 720–728. https://doi.org/10.2135/cropsci2011.07.0363

Lu, H., Price, J., Devkota, R., Rush, C., & Rudd, J. (2011). A dominant gene for resistance to wheat streak mosaic virus in winter wheat line CO960293-2. Crop Science, 51(1), 5–12. https://doi.org/10.2135/cropsci2010.01.0038

Lukaszewski, A. J., Alberti, A., Sharpe, A., Kilian, A., Stanca, A. M., Keller, B., Clavijo, B. J., Friebe, B., Gill, B., Wulff, B., Chapman, B., Steuernagel, B., Feuillet, C., Viseux, C., Pozniak, C., Rokhsar, D. S., Klassen, D., Edwards, D., Akhunov, E., … Feuillet, C. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science, 345(6194), 1251788–1251788. https://doi.org/10.1126/science.1251788

Malik, R., Brown-Guedira, G. L., Smith, C. M., Harvey, T. L., & Gill, B. S. (2003). Genetic mapping of wheat curl mite resistance genes *Cmc3* and *Cmc4* in common wheat. Crop Science, 43(2), 644–650. https://doi.org/10.2135/cropsci2003.6440

Martin, J. T., Fritz, A. K., Seifers, D. L., & Shroyer, J. P. (2007). 'RonL' hard white wheat. Kansas State University Agricultural Experiment Station and Cooperative Extension Service L-926, 3.

Martin, T. J., Zhang, G., Fritz, A. K., Miller, R., & Chen, M.-S. (2014). Registration of 'Clara CL' wheat. Journal of Plant Registrations, 8(1), 38–42. https://doi.org/10.3198/jpr2013.07.0040crc

McKelvy, U., Brelsford, M., Sherman, J., & Burrows, M. (2021). Reactions of winter wheat, spring wheat, and barley cultivars to mechanical inoculation with wheat streak mosaic virus. Plant Health Progress. https://doi.org/10.1094/PHP-10-20-0083-RS

Murugan, M., Cardona, P. S., Duraimurugan, P., Whitfield, A. E., Schneweis, D., Starkey, S., & Smith, C. M. (2011). Wheat curl mite resistance: Interactions of mite feeding with wheat streak mosaic virus infection. Journal of Economic Entomology, 104(4), 1406–1414. https://doi.org/10.1603/EC11112

Navia, D., de Mendonça, R. S., Skoracka, A., Szydło, W., Knihinicki, D., Hein, G. L., da Silva Pereira, P. R. V., Truol, G., & Lau, D. (2013). Wheat curl mite, Aceria tosichella, and transmitted viruses: An expanding pest complex affecting cereal crops. Experimental and Applied Acarology, 59(1–2), 95–143. https://doi.org/10.1007/s10493-012-9633-y

Oliveros, J. C. (2007). VENNY. An interactive tool for comparing lists with Venn Diagrams.

Park, C. J., & Seo, Y. S. (2015). Heat shock proteins: A review of the molecular chaperones for plant immunity. Plant Pathology Journal, 31(4), 323–333. https://doi.org/10.5423/PPJ.RW.08.2015.0150

Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R., & Nejentsev, S. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics, 28(21), 2747–2754. https://doi.org/10.1093/bioinformatics/bts526

Saintenac, C., Falque, M., Martin, O. C., Paux, E., Feuillet, C., & Sourdille, P. (2009). Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). Genetics, 181(2), 393–403. https://doi.org/10.1534/genetics.108.097469

Santra, M., Wang, H., Seifert, S., & Haley, S. (2017). Doubled haploid laboratory protocol for wheat using wheat–maize wide hybridization. In P. L. Bhalla & M. B. Singh (Eds.), Methods in Molecular Biology (Vol. 1679, pp. 235–249). Springer New York. https://doi.org/10.1007/978-1-4939-7337-8_14

See, D. R., Brooks, S., Nelson, J. C., Brown-Guedira, G., Friebe, B., & Gill, B. S. (2006). Gene evolution at the ends of wheat chromosomes. Proceedings of the National Academy of Sciences of the United States of America, 103(11), 4162–4167. https://doi.org/10.1073/pnas.0508942102

Seifers, D. L., Martin, T. J., Harvey, T. L., Fellers, J. P., & Michaud, J. P. (2009). Identification of the wheat curl mite as the vector of *Triticum* mosaic virus. Plant Disease, 93(1), 25–29. https://doi.org/10.1094/PDIS-93-1-0025

Seifers, D. L., Martin, T. J., Harvey, T. L., & Gill, B. S. (1995). Temperature sensitivity and efficacy of wheat streak mosaic virus resistance derived from *Agropyron intermedium*. Plant Disease, 79(11), 1104–1106. https://doi.org/10.1094/PD-79-1104

Seifers, D. L., Martin, T. J., Harvey, T. L., Haber, S., & Haley, S. D. (2006). Temperature sensitivity and efficacy of Wheat streak mosaic virus resistance derived from CO960293 wheat. Plant Disease, 90(5), 623–628. https://doi.org/10.1094/PD-90-0623

Seo, J.-K., & Kim, K.-H. (2016). Long-distance movement of viruses in plants. Current Research Topics in Plant Virology (pp. 153–172). Springer International Publishing. https://doi.org/10.1007/978-3-319-32919-2_6

Sharp, G. L., Martin, J. M., Lanning, S. P., Blake, N. K., Brey, C. W., Sivamani, E., Qu, R., & Talbert, L. E. (2002). Field evaluation of transgenic and classical sources of Wheat streak mosaic virus resistance. Crop Science, 42(1), 105–110. https://doi.org/10.2135/cropsci2002.1050

Singh, K., & Kundu, J. K. (2018). Wheat streak mosaic virus. Plant Viruses, 131–148. https://doi.org/10.1201/b22221-8

Singh, K., Wegulo, S. N., Skoracka, A., & Kundu, J. K. (2018). Wheat streak mosaic virus: A century old virus with rising importance worldwide. Molecular Plant Pathology, 19(9), 2193–2206. https://doi.org/10.1111/mpp.12683

Skoracka, A., Rector, B. G., & Hein, G. L. (2018). The interface between wheat and the wheat curl mite, *Aceria tosichella*, the primary vector of globally important viral diseases. Frontiers in Plant Science, 9(July), 1–8. https://doi.org/10.3389/fpls.2018.01098

Slykhuis, J. T. (1955). *Aceria tulipae Keifer* (Acarina: *Eriophyidae*) in relation to the spread of wheat streak mosaic. Phytopathology, 45(3), 116–128.

Stenger, D. C., Hall, J. S., Choi, I. R., & French, R. (1998). Phylogenetic relationships within the family *Potyviridae*: Wheat streak mosaic virus and brome streak mosaic virus are not members of the genus *Rymovirus*. Phytopathology, 88(8), 782–787. https://doi.org/10.1094/PHYTO.1998.88.8.782

Sun, Y., Qiao, Z., Muchero, W., & Chen, J. G. (2020). Lectin receptor-like kinases: The sensor and mediator at the plant cell surface. Frontiers in Plant Science, 11(December). https://doi.org/10.3389/fpls.2020.596301

Tan, C. T., Assanga, S., Zhang, G., Rudd, J. C., Haley, S. D., Xue, Q., Ibrahim, A., Bai, G., Zhang, X., Byrne, P., Fuentealba, M. P., & Liu, S. (2017). Development and validation of KASP markers for wheat streak mosaic virus resistance gene *Wsm2*. Crop Science, 57(1), 340–349. https://doi.org/10.2135/cropsci2016.04.0234

Tatineni, S., Alexander, J., Gupta, A. K., & French, R. (2019). Asymmetry in synergistic interaction between Wheat streak mosaic virus and *Triticum* mosaic virus in Wheat. Molecular Plant-Microbe Interactions, 32(3), 336–350. https://doi.org/10.1094/MPMI-07-18-0189-R

Tatineni, S., & French, R. (2014). The C-terminus of Wheat streak mosaic virus coat protein is involved in differential infection of wheat and maize through host-specific long-distance transport. Molecular Plant-Microbe Interactions, 27(2), 150–152. https://doi.org/10.1094/MPMI-09-13-0272-R

Tatineni, S., & Hein, G. L. (2018). Genetics and mechanisms underlying transmission of Wheat streak mosaic virus by the wheat curl mite. Current Opinion in Virology, 33, 47–54. https://doi.org/10.1016/j.coviro.2018.07.012

Tatineni, S., Kovacs, F., & French, R. (2014). Wheat streak mosaic virus infects systemically despite extensive coat protein deletions: Identification of virion assembly and cell-to-cell movement determinants. Journal of Virology, 88(2), 1366–1380. https://doi.org/10.1128/jvi.02737-13

Tatineni, S., Wosula, E. N., Bartels, M., Hein, G. L., & Graybosch, R. A. (2016). Temperature-dependent *Wsm1* and *Wsm2* gene-specific blockage of viral long-distance transport provides resistance to Wheat streak mosaic virus and Triticum mosaic virus in wheat. Molecular Plant-Microbe Interactions, 29(9), 724–738. https://doi.org/10.1094/MPMI-06-16-0110-R

Taylor, J., & Butler, D. (2017). R package ASMap: Efficient genetic linkage map construction and diagnosis. Journal of Statistical Software, 79. https://doi.org/10.18637/jss.v079.i06

Thind, A. K., Wicker, T., Šimková, H., Fossati, D., Moullet, O., Brabant, C., Vrána, J., Doležel, J., & Krattinger, S. G. (2017). Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. Nature Biotechnology, 35(8), 793–796. https://doi.org/10.1038/nbt.3877

Thomas, J. B., & Conner, R. L. (1986). Resistance to colonization by the wheat curl mite in *Aegilops squarrosa* and its inheritance after transfer to common wheat. Crop Science, 26(3), 527–530. https://doi.org/10.2135/cropsci1986.0011183x002600030019x

van Nocker, S., & Ludwig, P. (2003). The WD-repeat protein superfamily in *Arabidopsis*: Conservation and divergence in structure and function. BMC Genomics, 4(1), 50–50. https://doi.org/10.1186/1471-2164-4-50

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D., … Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature, 588(7837), 277–283. https://doi.org/10.1038/s41586-020-2961-x

Wells, D. G., Kota, R. S., Sandhu, H. S., Gardner, W. S., & Finney, K. F. (1982). Registration of one disomic substitution line and five translocation lines of winter wheat germplasm resistant to wheat streak mosaic virus (Reg. No. GP 199 to GP 204). Crop Science, 22(6).

Whelan, E. D. P., & Hart, G. E. (1988). A spontaneous translocation that transfers wheat curl mite resistance from decaploid *Agropyron elongatum* to common wheat. Genome, 30(3), 289–292. https://doi.org/10.1139/g88-050

Whitham, S. A., Anderberg, R. J., Chisholm, S. T., & Carrington, J. C. (2000). *Arabidopsis RTM2* gene is necessary for specific restriction of tobacco etch virus and encodes an unusual small heat shock–like protein. The Plant Cell, 12(4), 569–582. https://doi.org/10.1105/tpc.12.4.569

Xie, Y., Ravet, K., & Pearce, S. (2021). Extensive structural variation in the Bowman-Birk inhibitor family in common wheat (*Triticum aestivum* L.). BMC Genomics, 22(1), 218–218. https://doi.org/10.1186/s12864-021-07475-8

Xu, J., Wang, X. Y., & Guo, W. Z. (2015). The cytochrome P450 superfamily: Key players in plant development and defense. Journal of Integrative Agriculture, 14(9), 1673–1686. https://doi.org/10.1016/S2095-3119(14)60980-1

Zhang, G., & Hua, Z. (2018). Genome comparison implies the role of *Wsm2* in membrane trafficking and protein degradation. PeerJ, 2018(4), 1–18. https://doi.org/10.7717/peerj.4678

Zhang, G., Martin, T. J., Fritz, A. K., Miller, R., Chen, M.-S., Bowden, R. L., & Bai, G. (2016). Registration of 'Joe' hard white winter wheat. Journal of Plant Registrations, 10(3), 283–286. https://doi.org/10.3198/jpr2016.02.0007crc

Zhang, G., Martin, T. J., Fritz, A. K., Miller, R., Chen, M.-S., Bowden, R. L., & Johnson, J. J. (2015). Registration of 'Oakley CL' wheat. Journal of Plant Registrations, 9(2), 190–195. https://doi.org/10.3198/jpr2014.04.0023crc

Zhu, T., Wang, L., Rimbert, H., Rodriguez, J. C., Deal, K. R., De Oliveira, R., Choulet, F., Keeble-Gagnère, G., Tibbits, J., Rogers, J., Eversole, K., Appels, R., Gu, Y. Q., Mascher, M., Dvorak, J., & Luo, M.-C. (2021). Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. The Plant Journal, 107(1), 303–314. https://doi.org/10.1111/tpj.15289

CHAPTER 4. TRANSCRIPTOMICS OF BANANA (*MUSA ACCUMINATA*) IN RESPONSE TO *FUSARIUM OXYSPORUM* F.SP. *CUBENSE* (*FOC*) SUBTROPICAL RACE 4 (STR4) INFECTION

4.1 Summary

Banana (*Musa accuminata*) is an important staple food in the developing world and one of the leading fruit crops globally. One of the main diseases limiting its production is *Fusarium* wilt disease caused by the fungal pathogen *Fusarium oxysporum* f.sp. *cubense* (*Foc*). The *Foc* race 4 strain, first detected 50 years ago in South East Asia, has recently spread globally and is virulent to most banana cultivars on the market, making it urgent to identify genetic resistance and mitigate its effect on global banana production. To characterize the host transcriptomic response to subtropical race 4 (*Foc*-STR4) infection, our collaborators generated RNA-seq data from resistant and susceptible samples prior to infection (T0), and 1-, 3-, and 7-day post inoculation (T1, T3, and T7) with *Foc*-STR4. Time course transcriptomic analysis of samples indicated host plant undergo major transcriptional reprogramming after *Foc* infection, including the immediate broad down-regulation of signaling transduction and biosynthetic related genes. Comparison of genes differentially expressed between resistant and susceptible genotypes revealed that common defense responses, such as reactive oxygen species (ROS) production and cell wall modification, occur in both resistant and susceptible materials after *Foc* infection. However, these responses were much slower in susceptible cultivars compared to resistant cultivars, indicating plant resistance is likely achieved by immediate induced signaling transduction as well as rapid defense responses. Our collaborators performed a QTL-seq study and identified a novel locus that confers resistance to *Foc*-STR4. Analysis of candidates underlying the locus identified thirteen genes that

were differentially expressed between genotypes, the full description of these candidates will be published alongside the QTL-seq study.

4.2 Introduction

Banana is one of the most important fruits in the world with over 100 million tons produced annually (FAOSTAT, 2020). Global banana production is severely affected by *Fusarium* wilt disease, also known as Panama disease, which is especially devastating in Asia, Australia, the Middle East, and Africa (Ploetz, 2015). In the Philippines alone, economic losses can reach $400 million per year (Cook et al., 2015). This disease is caused by the soil-born hemi-biotrophic fungal pathogen *Fusarium oxysporum* f.sp. *cubense* (*Foc*) (Ploetz, 2006). *Foc* initially infects plants by attaching to host root hairs. Once inside the root, *Foc* can colonize the rhizomes and then move to the pseudostem, blocking the xylem vessels and eventually preventing water and nutrient transport (Li et al., 2013). Disease symptoms of *Fusarium* wilt include yellowing and wilting of leaf tissues, and brown discoloration and necrosis of xylem vessels in the rhizomes and stems (Ploetz, 2015).

The banana cultivar ‘Cavendish’ provides 99% of bananas grown for export, and exhibits resistance to most historical *Foc* races (race 1, race 2, and race 3), mitigating the impact of *Fusarium* wilt disease (Ploetz, 2006). However, a new *Foc* strain named as race 4, evolved around 1970 in South East Asia, is virulent for ‘Cavendish’ and many other banana cultivars (Ploetz, 1994). The race 4 strain is further divided into tropical race 4 (*Foc*-TR4) and subtropical race 4 (*Foc*-STR4), according to their infection areas (Buddenhagen, 2009). The initial outbreak of race 4 *Foc* leads to spread worldwide so that this strain now presents a major threat to global banana production (Ploetz, 2015). Chemical controls against *Foc*, such as soil fumigants, fungicides, and cultural practices are uneconomic, ineffective, and environmentally unfriendly (Sismak & Zheng,

2018). The long-term solution to manage this disease is to identify sources of genetic resistance and introduce the resistance genes into commercial cultivars.

Most cultivated bananas have a triploid genome (2n = 3x = 33, genome constitutions of AAA, AAB, or ABB), derived from two diploid progenitors, *Musa acuminata* (AA genome) and *Musa balbisiana* (BB genome) (D'Hont et al., 2000). There is a lack of natural disease-resistant germplasm for race 4 *Foc* strain among commercial banana cultivars (Chen et al., 2019). Efforts have been made to screen banana wild relatives to identify resistance sources against the virulent race 4 *Foc* strain. A wild banana accession 'Pahang' (*Musa acuminata* ssp. *malaccensis*, n = 11, A genome), was identified with strong resistance to *Foc*-TR4 (D'hont et al., 2012; Zhang et al., 2018). Evaluation of 'Pahang' revealed the resistance mechanism is through suppression of fungal growth in the corm (Zhang et al., 2018). Moreover, the first banana reference genome assembly is from a doubled haploid 'Pahang' (DH-Pahang, AA genome), with a genome size of 523 Megabase pairs (Mbp) containing 36,542 protein-coding gene models (D'hont et al., 2012).

Additionally, a screen of 129 diverse banana accessions identified ten that exhibit strong host resistance to *Foc*-TR4 (Zuo et al., 2018). Moreover, the potential 'complete' resistance in the rhizomes has been identified from a recent screen of 34 banana cultivars grown under controlled settings, providing valuable genetic resources for *Fusarium* wilt disease management (Chen et al., 2019). To deploy such genetic resources into elite banana cultivars, genetic studies are required to map the resistance loci and identify causative genes. In addition, a better understanding of host responses to infection can help define plant defense mechanisms.

A team at the University of Queensland has been working on the identification of QTLs controlling host resistance to *Foc* race 4 in bananas. They performed fine mapping using 430 F_2 individuals and identified a novel *Foc* race 4 type resistance locus in a 4.3 cM genetic interval on

chromosome 3. In this study, I analyzed the gene expression profiles of these candidate genes between resistant and susceptible cultivars in a time course prior to infection (T0), and 1-, 3-, and 7-day (T1, T3, and T7) after infection with *Foc*-STR4. Moreover, I assembled gene co-expression networks for susceptible samples to study transcriptional changes in overall plant defense responses during infection. This analysis revealed that although susceptible samples still respond to *Foc*-STR4 infection by inducing ROS production and cell wall strengthening related genes, these defense responses were induced more slowly than in resistant materials. Moreover, comparing resistant *versus* susceptible samples for DEGs underlying the novel *Foc*-STR4 locus helped prioritized thirteen candidates that will be subjected to future functional validation assay.

4.3 Materials and Methods

The details of data preparation (plant materials, inoculation approach, pathogen strains, RNA-seq library prep etc.) were performed by our collaborators and will be described in full in an upcoming publication.

4.3.1 Experimental design

The RNA-seq experiment comprised 24 samples from *Musa acuminata* ssp. *malaccensis* with two genotypes: resistant and susceptible; four time points: 0-, 1-, 3-, and 7-day post inoculation (dpi), where 0 dpi is prior to *Foc* infection; and three biological replicates (2 * 4 * 3 = 24 samples, Table S4.1). Each cDNA library was sequenced using the Hiseq 4000 platform (Genewiz), generating approximately 150 bp paired end reads for each sample.

4.3.2 Differentially expressed genes (DEGs) analysis

Reads containing adapter sequence and low-quality reads were removed using Fastp software (Chen et al., 2018). The paired-end filtered reads were aligned to an unpublished banana reference genome based on DH-Pahang shared by our collaborators. Differentially expressed genes (DEGs)

were analyzed using two different pipelines. The first pipeline used STAR for alignment (Dobin & Gingeras, 2015) with parameters "-outFilterMismatchNmax 6 -alignIntronMax 10000". Non-normalized reads were tabulated with FeatureCounts software (option: -M -g ID -t gene -p) (Liao et al., 2014). DEGs were identified from pairwise comparisons between resistant and susceptible samples at each time point ($T0_S^R$, $T1_S^R$, $T3_S^R$, and $T7_S^R$) and comparing *Foc* treated samples versus untreated samples (T0 *vs.* T1, T0 *vs.* T3, and T0 *vs*. T7) using DESeq2 R package (Love et al., 2014). The *P*-value threshold was determined using Benjamini and Hochberg's approach (Benjamini & Hochberg, 1995) for controlling the false discovery rate (FDR < 0.01) without controlling the $\log_2$ fold change (FC).

The second pipeline used HISAT2 v2.1.0 (Kim et al., 2019) with default settings for alignment of filtered reads. The alignments were subjected to samtools v1.11 (Danecek et al., 2021) to generate sorted BAM files, which were used as input for StringTie v2.0.3 (Pertea et al., 2015) to perform *de novo* transcriptome assembly to identify novel splice variants. Outputs from StringTie were imported into the Ballgown R package (Frazee et al., 2015) for statistical tests to identify DEGs between genotypes based on gene- and transcript-level fragments per kilobase of transcript per million fragments mapped (FPKM) values (Pertea et al., 2016). The default statistical tests in Ballgown were used, based on a parametric F-test comparing nested linear models (Frazee et al., 2015). Multiple comparison adjustments were reported by q-values (Storey & Tibshirani, 2003) for each transcript, and DEGs were defined as those genes with a q-value less than 0.01. Gene Ontology (GO) enrichment analysis was performed with the topGO R package (Alexa A, 2021) running Fisher test for significantly enriched GO terms ($P < 0.01$ for DEGs from modules identified by WGCNA, and $P < 0.05$ for DEGs between genotypes) for biological process (BP), molecular function (MF), and cellular component (CC). The enriched GO terms together with its

Fisher test values were selected as inputs to remove redundant GO terms on Revigo webpage (Supek et al., 2011) with the parameter: “tiny-0.4”.

4.3.3 Gene co-expression network assembly (WGCNA) analysis

Both DESeq2 and Impulse2 (Spies et al., 2019) were used to call DEGs between time points (P adj < 0.01) and throughout the time course (T0, T1, T3, and T7). Comparing both statistical tests, shared DEGs from both tests were subjected to WGCNA analysis to construct a gene co-expression network. A customized R script (wgcna_standard.R, Appendix C.1) adapted from a standard WGCNA analysis pipeline (Langfelder & Horvath, 2008) was used with the parameters power = 20, minmoduleSize = 30.

4.4 Results

4.4.1 Host plants undergo major transcriptomic reprogramming after *Foc* infection

An average of 48.3 million clean 150-bp paired end (PE) reads were generated for 24 RNA-seq samples after filtering (Table S4.2). The clean reads were aligned to a banana reference genome using two pipelines to compare the performance of STAR and HISAT2 alignment tools. Using STAR, the average overall mapping rate was 95.6% $\pm$ 1.6%, while the mean unique mapping rate was 87.8% $\pm$ 5.1% (Table S4.2). In contrast, using HISAT2 the average overall mapping rate was 96.5% $\pm$ 0.9%, and the average unique mapping rate was 82.8% $\pm$ 8.1% (Table S4.2). Because the STAR alignment pipeline yields a higher unique mapping rate than HISAT2, the outputs from STAR were used for all subsequent analyses.

A principal component analysis (PCA) plot and sample-to-sample distance matrix were built based on whole transcriptome data from each sample (Figure 4.1). In general, samples were grouped distinctly by timepoint with little overlap between them, showing time points are the major driver of transcriptional differences in these samples. In the PCA plot, PC1 explained 72%

of variance in the overall transcriptome between samples, and most clearly distinguished samples prior to *Foc* inoculation (T0) from samples after *Foc* inoculation (T1, T3, and T7, Figure 4.1A). This profile indicates that both susceptible and resistant host plants undergo major transcriptomic reprogramming soon after *Foc* infection. Moreover, PC2 explained 11% of variance, separating samples T1, T3, and T7 from one another (Figure 4.1A), suggesting host plants underwent relatively smaller transcriptional changes from 24 hours after the initial infection. Despite major differences in transcription between timepoints, there were relatively smaller differences between genotypes, which remained clustered together by timepoint (Figure 4.1B).

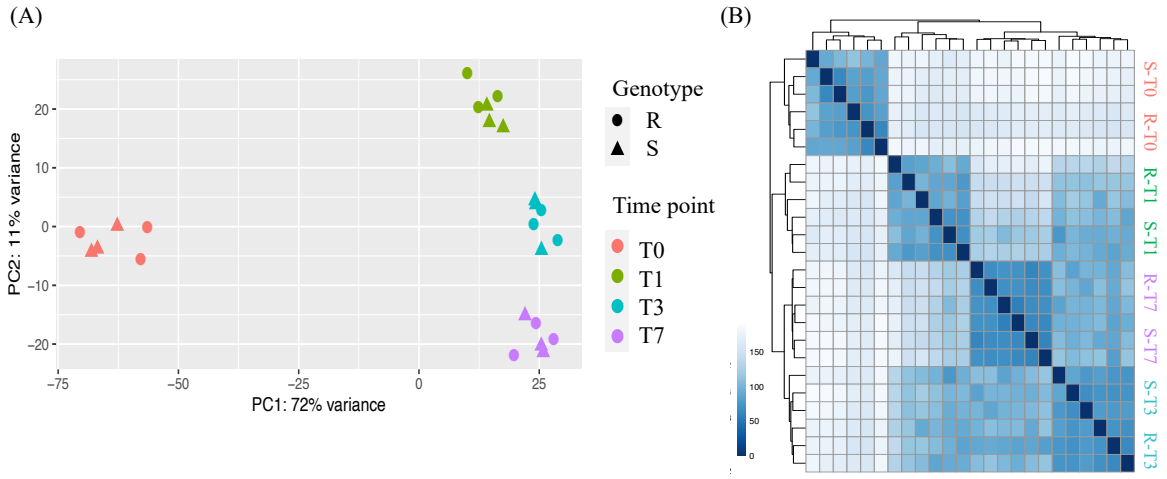


Figure 4.1. Summary of RNA-seq samples. **(A)** Principal component plot (PCA). The unit indicated percent variant explained (PVE). RNA-seq samples include three biological replicates for two genotypes (Resistant-R, and Susceptible-S) and four time points (before inoculation-T0, 1-, 3-, and 7-day post inoculation- T1, T3, and T7). **(B)** Sample correlation matrix. The heatmap shows the distance matrix for similarities and dissimilarities in expression profiles between samples.

4.4.2 Co-expression networks reveal major host plant transcriptomic profiles following *Foc* infection

To characterize the host transcriptomic response to *Foc* infection over time in susceptible genotypes, pairwise DEG analysis was performed between different stages following *Foc* infection and untreated T0 samples (T0 *vs.* T1, T0 *vs.* T3, and T0 *vs*. T7). The banana reference genome includes 36,443 gene models, of which 32,336 (88.7%) were expressed with ≥ 10 read counts in

at least one of the samples. In response to *Foc* infection, more than half of all expressed genes (16,278/32,336 = 50.3%) were differentially expressed (DESeq2 *P* adj < 0.01) in at least one pairwise comparison between treated and untreated samples, consistent with the PCA plot showing major differences between untreated (T0) and infected samples (T1, T3, and T7). At each pairwise comparison between T0 with time points after *Foc* infection (T1, T3, and T7), there were a greater number of genes upregulated than downregulated (Figure 4.2A). Out of the three pairwise comparisons, 5,637 DEGs were shared (Figure 4.2B), indicating those genes were induced or repressed throughout the time course from T1 to T7. In addition, 1,770, 1,577, and 3,266 genes were differentially expressed at only one time point (T1, T3, and T7) compared to T0, respectively (Figure 4.2B), indicating that temporary or stage specific transcriptional changes are greatest at the later time point, seven days after *Foc* infection.

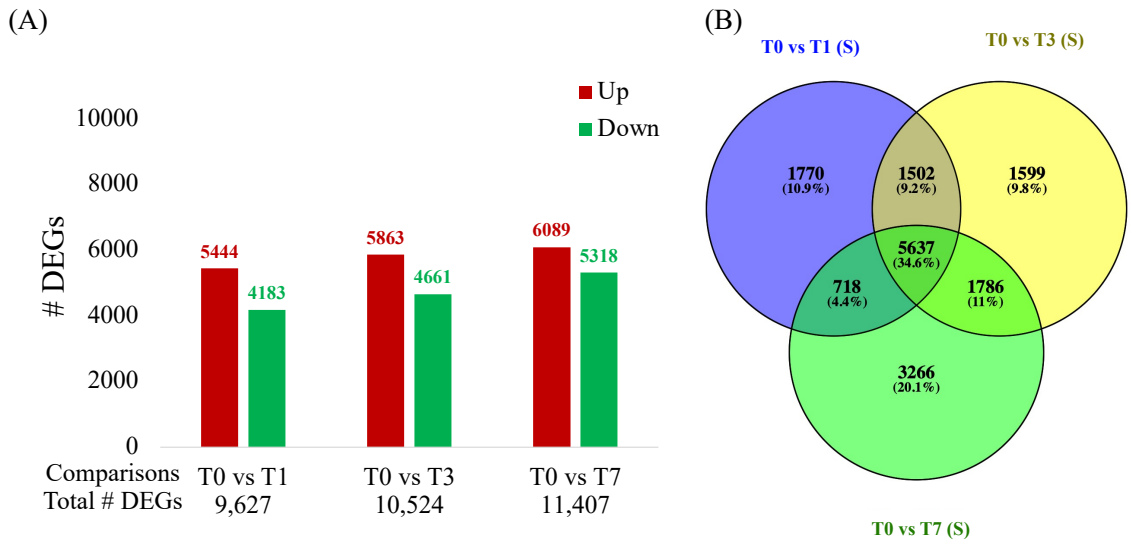


Figure 4.2. Summary of differentially expressed genes (DEGs) from pairwise comparisons of T1, T3, and T7 *versus* T0 in susceptible samples. **(A)** Bar chart for number of up- and down-regulated DEGs T0 *vs.* T1, T0 *vs.* T3, and T0 *vs.* T7; **(B)** Venn diagram for total number of DEGs in each comparison. T0 means prior to *Foc* treatment, T1, T3, and T7 means 1-, 3-, and 7-day post *Foc* inoculation (dpi).

Using Impulse2, a tool to detect DEGs across time course expression data, 13,100 DEGs were identified, 12,961 of which were also significant by DESeq2 pairwise comparisons. This consensus set of DEGs was used to construct a co-expression network to characterize the predominant expression profiles during the infection time course. The network includes 12 modules, with the largest number of genes (3,716) in Module 1 (Table S4.3). Genes in this module are characterized by high transcript levels immediately before infection, followed by rapid downregulation and suppression following *Foc* infection for the remainder of the time course (Figure 4.3). Genes in Module 1 were most significantly enriched for the gene molecular function (MF) terms "DNA-binding transcription factor activity" (GO:0003700), and for the biological processes (BP) terms "regulation of salicylic acid mediated signaling pathway" (GO:2000031), "regulation of signaling" (GO:0023051)", "regulation of cell communication" (GO:0010646), "regulation of response to stimulus" (GO:0048583), "biosynthetic process" (GO:0045927), and "regulation of cellular process" (GO:0050794) (Table S4.4). These broad profiles indicate that following *Foc* infection, susceptible host plants rapidly suppress the expression of large groups of TFs that may regulate signal transduction, cellular process, and other biosynthesis processes. Furthermore, the downregulation of salicylic acid (SA) signaling pathways indicates that SA might plays a role in host plant response to this hemi-biotrophic pathogen's infection.

The genes clustered in Module 3 (1,812 DEGs) and Module 4 (1,339 DEGs) were induced rapidly at T1 and were significantly enriched for the BP terms "protein phosphorylation" (GO:0006468), "response to abiotic stimulus" (GO:0009628), "response to cold" (GO:0009409), and "response to red or far-red light" (GO:0009639) (Table S4.4), indicating following *Foc* infection host plant induced transcriptional changes to abiotic stress within 24 hours.

In contrast, the genes in the second largest module (Module 2, 2,553 DEGs) exhibited low transcript levels at T0 and T1 followed by rapid upregulation at T3 which is maintained throughout the remainder of the time course (Figure 4.3). Enriched GO terms for BP in Module 2 included hydrogen peroxide activity or reactive oxygen species (ROS) production (GO:0006979, GO:0042744, GO:0072593), and cell wall/lignin biogenesis (GO:0042546, GO:0046274) related terms (Table S4.4). The cell wall biogenesis and production of ROS related biological processes indicate host plants may activate these defense responses to fight against *Foc*, but responses were induced only 72 hours after infection.

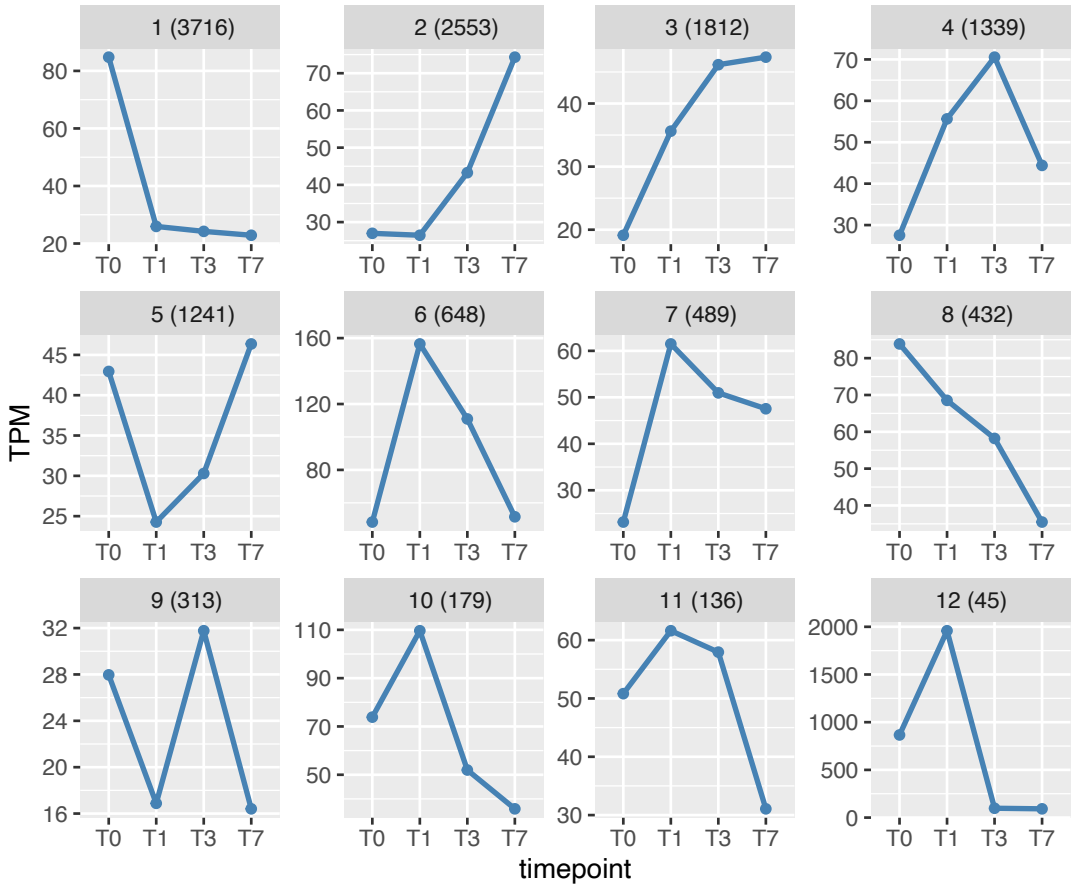


Figure 4.3. Expression profiles for modules identified from the WGCNA analysis. The expression for each module were displayed as the average TPM values of the eigengene representing that module. The module number and the number of DEGs in that module (in parenthesis) are displayed above the expression line and highlight in grey color.

Collectively, the WGCNA result suggested that following *Foc*-STR4 infection, host plants undergo transcriptional changes to abiotic stress response, as well as repressed SA signaling and other cell communication and biosynthetic processes within 24 hours. The biotic stress induced process, such as ROS production and cell wall strengthening were induced until 72 hours.

4.4.3 Host defense related response was different between genotypes 24 hours after *Foc* infection

In pairwise comparisons between resistant and susceptible genotypes (R *vs.* S), 1,097 genes (3.4% of all expressed genes), were differentially expressed in at least one timepoint (Table S4.5), consistent with the PCA plot result showing that there are comparatively fewer genes differentially expressed between genotypes than between time points. GO enrichment tests were performed separately for DEGs that were more highly expressed in resistant or susceptible materials at each time point (T0, T1, T3, and T7) (Table S4.6).

Time points	Total #DEGs	Higher in Resistant	Higher in Susceptible
T0	316	140	176
T1	686	113	573
T3	58	26	32
T7	37	17	20
Total	1,097	296	801

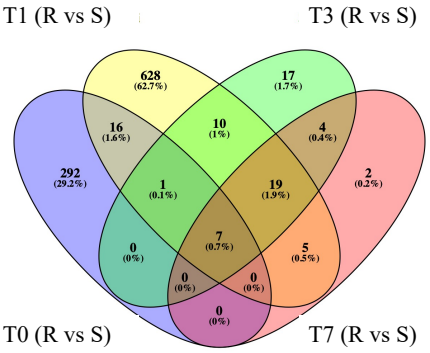


Figure 4.4. Summary of DEGs between resistant versus susceptible samples (R *vs.* S). The Venn diagram shows the overlapping DEGs at each time point and the table below shows total number of DEGs at each time point. "Higher in Resistant" means these DEGs were upregulated in resistant sample and "Higher in Susceptible" means DEGs were downregulated in resistant samples

At T0 before infection, 316 DEGs were identified, of which 140 DEGs were more highly expressed in resistant samples compared to 176 that were more highly expressed in susceptible samples (Figure 4.4). The great majority of these genes (292, or 92.4%) were differentially

expressed only at T0 (Figure 4.4), indicating they likely correspond to genotypic variation under unstressed conditions and may play a role in preparing plants for resistance to *Foc* infection.

The majority of DEGs between genotypes were found at T1, which has 686 (686/1,097 = 63%) DEGs, of which 113 had higher expression levels in resistant plants and 573 had higher expression levels in susceptible plants (Figure 4.4). Among them, there were 628 DEGs unique to this time point (Figure 4.4), indicating under stressed conditions the genes causing main genotypic difference for transcriptional changes to *Foc* infection happened within 24 hours. In comparison, fewer transcriptomic differences between genotypes were detected at later timepoints, with just 58 DEGs identified at T3 and 37 DEGs identified at T7 (Figure 4.4). At T1, genes more highly expressed in resistant plants were enriched for many defense related GO terms such as "immune system process" (GO:0002376), "response to stress" (GO:0006950), "cell wall biogenesis" (GO:0042546), "response to oxidative stress" (GO:0006979), and "defense response to other organism" (GO:0098542) (Table S4.6). The result suggested that resistant plants likely activate these rapid defense responses within 24 hours after infection.

4.4.4 Expression profiles of candidate genes underlying a STR4 resistance locus

Our collaborators performed a QTL-seq study and identified a novel locus associated with resistance to *Foc*-STR4 in an interval of approximately 400 Kbp. Annotations for candidate genes underlying this QTL revealed a cluster of adjacent genes belonging to the same gene family. In total, 59 genes belong to this gene family were identified within the QTL interval, and can be grouped into four subfamilies, A, B, C and D, named according to the order of their physical position on the banana reference genome (Table S4.7). In total, there are 41 gene members in subfamily A (named A-1, A-2 etc.), nine genes in subfamily B, three genes in subfamily C, and six genes in subfamily D (Table 4.7). The expression profiles of these genes across the time course

were quantified as mean transcript per million (TPM) values, and pairwise DEG analysis was performed to compare expression levels between resistant and susceptible genotypes (Table S4.7).

Most members of this family exhibit low transcript levels and only eight genes exhibited a mean expression level > 1 TPM (Figure 4.5A). Among the four expressed genes in subfamily A, two (*A-1* and *A-36*) were only expressed in the resistant genotype, whereas the other two (*A-5* and *A-29*) only expressed in susceptible genotype (Figure 4.5A). Moreover, two genes in subfamily C (*C-1* and *C-3*) and one gene in subfamily D (*D-5*) were highly expressed in both resistant and susceptible genotypes throughout the time course, while another candidate gene (*C-2*) was expressed only at T3 and T7 (Figure 4.5A). Moreover, thirteen gene*s* in this family were differentially expressed between genotypes ($P < 0.001$) in at least one time point (Figure 4.5B). Nine of these genes in subfamilies A and B were more highly expressed in resistant plants, while the remaining four genes from subfamilies C and D were more highly expressed in susceptible plants (Figure 4.5B).

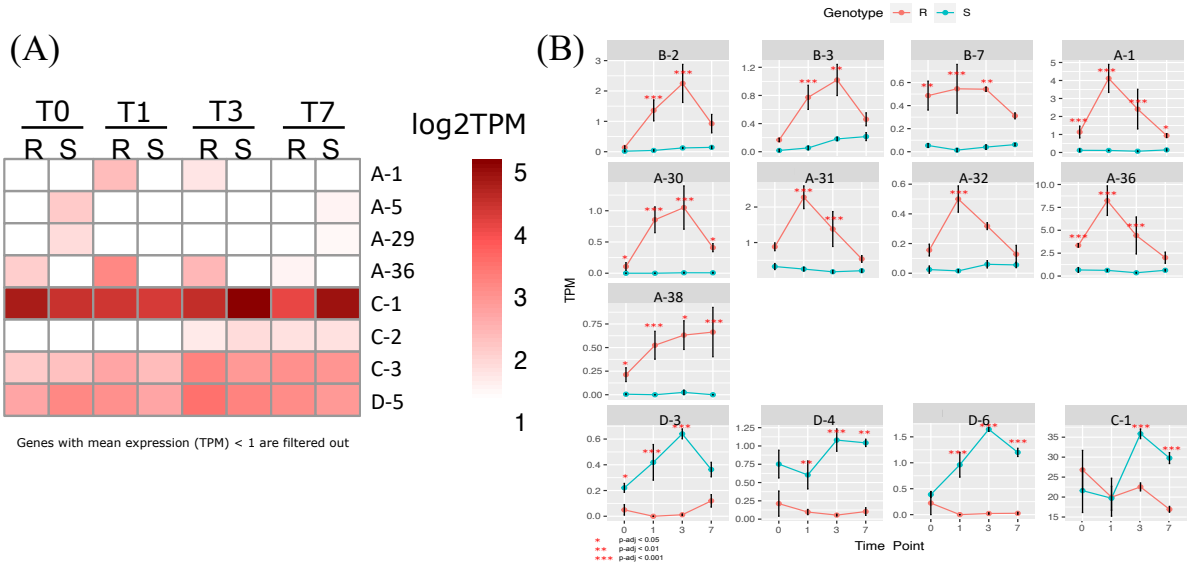


Figure 4.5. Expression profiles of candidates using STAR as the alignment tool and DESeq2 for statistical test to identify DEGs between resistant versus susceptible samples. **(A)** Heatmap of candidate genes after filtering out genes with average TPM < 1 across samples. Heatmap was generated based on log_2TPM values. **(B)** Line plot of candidate genes that were significantly expressed between genotype ($P < 0.001$), the star indicates the significance level between genotype at each time point.

4.5 Discussion

In this study, transcriptomic data from banana root tissue was used to characterize host response to *Foc* STR4 strain infection. Overall, major transcriptional reprogramming for both resistant and susceptible genotypes occurred in response to *Foc* infection compared to uninoculated controls (Figure 4.1A). More than 50% of transcribed genes were differentially expressed in susceptible samples following *Foc* infection (Figure 4.2A). These findings are consistent with earlier studies of *Foc* infection. For example, Wang et al., (2012) identified large transcriptomic changes induced by *Foc*-TR4 in susceptible banana roots, with 4,729, 5,078, and 5,531 DEGs 2, 4, and 6 dpi, respectively, using a *de novo* assembled banana reference genome that consisted of 21,622 gene models. Likewise, Sun et al., (2019) identified 9,612 DEGs in both *Foc*-TR4 resistant and susceptible materials comparing different infection stages (2, 4, and 6 dpi) to 0 dpi, confirming transcriptional reprogramming of banana roots in response to *Foc*-TR4 infection. However, this is inconsistent with a study on the susceptible cultivar 'Cavendish' in response to *Foc*-TR4, where just 473, 722, and 1,043 DEGs were identified 3, 27, and 51 hours after *Foc*-TR4 infection in root tissues (Li et al., 2013). One reason for these differences could be the study design, such that in 'Cavendish', mock inoculated samples were used as the control group, whereas in the current study the T0 was uninoculated samples. The DEGs detected in our study comparing T1, T3, and T7 *versus* T0 may include transcriptional changes induced by the plant's response to wounding or other mechanical stresses challenged with inoculation.

Gene co-expression networks in susceptible materials revealed the broad transcriptomic profiles and changes in host molecular processes following *Foc* STR4 infection. Signaling transduction, response to stimulus, and SA-mediated signaling pathways related genes were downregulated and suppressed in susceptible samples within 24 hours of *Foc* infection (Figure 4.3

and Table S4.4). Signal transduction pathways play an indispensable role in activating plant defense responses after the perception of pathogens and these results suggest that repression in signal transduction pathways may be associated with plant susceptibility to *Foc*. Among the plant hormone-related signaling pathways known to regulate defense gene expression, jasmonic acid (JA) and ethylene (ET) pathways contribute to host resistance against *Foc* race 4, whereas the role of salicylic acid (SA) is not known yet (Li et al., 2012; Swarupa et al., 2014). Our study suggested that the SA pathways were suppressed in susceptible plants immediately after infection and may be a contributing factor to plant susceptibility to *Foc*-STR4.

In addition to repressed signaling transduction, we found susceptible host plants induce other defense responses, such as ROS production and cell wall biogenesis (Table S4.4), although not until 72 hours after infection (Figure 4.3). Plants have developed a series of defense responses to fight against pathogen attack, and immediately after recognition of pathogen signals one common response is the production of ROS (Swarupa et al., 2014). This is consistent with Wang et al., (2012) who found that ROS production pathways were enriched in susceptible banana roots following infection with *Foc*-TR4. They concluded that banana roots responded to infection by *Foc*-TR4 through ROS production, but that these early defense activities in susceptible genotypes is not sufficient to provide resistance against the pathogen (Wang et al., 2012). Other than ROS production, cell wall strengthening is another important plant defense response against fungal pathogens that is usually induced in both susceptible and resistant materials (Bai et al., 2013; Swarupa et al., 2014).

This study also investigated transcriptional changes between resistant and susceptible genotypes after *Foc* STR4 infection. Of the 1,097 genes differentially expressed between resistant and susceptible genotypes (Figure 4.4), the majority were identified within 24 hours after infection,

and were highly enriched for ROS production, cell wall biogenesis, and other stress related plant defense responses (Table S4.6). This result indicated that while these defense responses were induced in both genotypes, the timing and abundance is different between resistant and susceptible genotypes, and that the resistant genotypes have a much faster defense response. These findings were also reported in Bai et al., (2013) that demonstrated much faster early immune response in resistant cultivar compared to susceptible cultivar in response to *Foc*-TR4.

Plants have two layers of innate immunity against pathogens, characterized by transmembrane pattern recognition receptors (PRRs) recognition of pathogen or microbial associated molecular patterns (PAMPS/MAMPs) triggered immunity (PTI), and intracellular nucleotide-binding site leucine-rich repeat receptor (NLR) type of resistance (*R*) gene perception of pathogen effector triggered immunity (ETI) (Jones & Dangl, 2006). Comparing the two immunity responses, there is substantial overlap between downstream defense responses during PTI and ETI after the initial perception of pathogens. Both responses include activation of mitogen-activated protein kinases (MAPKs), oxidative burst, and ion influx, suggesting the defense signaling converges in PTI and ETI (Navarro et al., 2004). However, compared to the prolonged and higher magnitude of ROS production and MAPK activation in ETI, the downstream signaling triggered in PTI is much more rapid and transient (Cui et al., 2015; Tsuda & Katagiri, 2010). Considering differences in the magnitude and duration of early oxidative burst and cell wall lignification related defense responses that differentiate *Foc*-STR4 resistance from susceptibility in this study, the resistance mechanism is likely related to the initial recognition of the pathogen that leads to PTI responses. Further studies will be needed to characterize genes that involved in PTI signaling transduction and clarify plant immunity mechanism against *Foc*-STR4.

Outbreaks of *Foc* race 4 strain due to its high virulence and wide host range poses a major threat to global banana production, making it urgent to look for host genetic resistance to control this disease (Ploetz, 1994). Our collaborators conducted a QTL-seq study and identified a novel locus with resistance to *Foc*-STR4. To further narrow candidate genes underlying the locus, I identified thirteen genes differentially expressed between resistant and susceptible genotypes (Figure 4.5). These thirteen genes are considered top candidates for functional validation with CRISPR/Cas9, and a detailed description of candidate genes and their characterization will be published as part of the QTL-seq study.

REFERENCES

Alexa A, R. J. (2021). TopGO: Enrichment Analysis for Gene Ontology.

Bai, T.-T., Xie, W.-B., Zhou, P.-P., Wu, Z.-L., Xiao, W.-C., Zhou, L., Sun, J., Ruan, X.-L., & Li, H.-P. (2013). Transcriptome and expression profile analysis of highly resistant and susceptible banana roots challenged with *Fusarium oxysporum* f. Sp. *cubense* Tropical Race 4. PLOS ONE, 8(9), e73945.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Buddenhagen, I. (2009). Understanding strain diversity in *Fusarium oxysporum* f. Sp. *cubense* and history of introduction of 'Tropical race 4' to better manage banana production. Acta Horticulturae, 828, 193–204. https://doi.org/10.17660/ActaHortic.2009.828.19

Chen, A., Sun, J., Matthews, A., Armas-Egas, L., Chen, N., Hamill, S., Mintoff, S., Tran-Nguyen, L. T. T., Batley, J., & Aitken, E. A. B. (2019). Assessing variations in host resistance to *Fusarium oxysporum* f sp. *cubense* race 4 in *Musa* species, with a focus on the subtropical race 4. Frontiers in Microbiology, 10(MAY). https://doi.org/10.3389/fmicb.2019.01062

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cook, D. C., Taylor, A. S., Meldrum, R. A., & Drenth, A. (2015). Potential economic impact of Panama disease (tropical race 4) on the Australian banana industry. Journal of Plant Diseases and Protection, 122(5/6), 229–237.

Cui, H., Tsuda, K., & Parker, J. E. (2015). Effector-triggered immunity: From pathogen perception to robust defense. Annual Review of Plant Biology, 66(1), 487–511. https://doi.org/10.1146/annurev-arplant-050213-040012

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience, 10(2). https://doi.org/10.1093/gigascience/giab008

D'hont, A., Denoeud, F., Aury, J. M., Baurens, F. C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., Da Silva, C., Jabbari, K., Cardi, C., Poulain, J., Souquet, M., Labadie, K., Jourda, C., Lengellé, J., Rodier-Goud, M., … Wincker, P. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature, 488(7410), 213–217. https://doi.org/10.1038/nature11241

D'Hont, A., Paget-Goy, A., Escoute, J., & Carreel, F. (2000). The interspecific genome structure of cultivated banana, *Musa* spp. Revealed by genomic DNA in situ hybridization. Theoretical and Applied Genetics, 100(2), 177–183. https://doi.org/10.1007/s001220050024

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. Current Protocols in Bioinformatics, 51(1), 11.14.1-11.14.19. https://doi.org/10.1002/0471250953.bi1114s51

FAOSTAT. (2020). Food and Agriculture Organization of the United Nations Database, FAOSTAT statistics, Crops-FAOSTAT statistics, Crops. https://doi.org/10.1016/B978-0-12-384947-2.00270-1

Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., & Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. Nature Biotechnology, 33(3), 243–246. https://doi.org/10.1038/nbt.3172

Jones, J. D. G., & Dangl, J. L. (2006). The plant immune system. Nature, 444(7117), 323–329. https://doi.org/10.1038/nature05286

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1), 559–559. https://doi.org/10.1186/1471-2105-9-559

Li, C., Shao, J., Wang, Y., Li, W., Guo, D., Yan, B., Xia, Y., & Peng, M. (2013). Analysis of banana transcriptome and global gene expression profiles in banana roots in response to infection by race 1 and tropical race 4 of *Fusarium oxysporum* f. Sp. *cubense*. BMC Genomics, 14(1), 851–851. https://doi.org/10.1186/1471-2164-14-851

Li, C. yu, Deng, G. ming, Yang, J., Viljoen, A., Jin, Y., Kuang, R. bin, Zuo, C. wu, Lv, Z. cheng, Yang, Q. song, Sheng, O., Wei, Y. rong, Hu, C. hua, Dong, T., & Yi, G. jun. (2012). Transcriptome profiling of resistant and susceptible Cavendish banana roots following inoculation with *Fusarium oxysporum* f. Sp. *cubense* tropical race 4. BMC Genomics, 13(1). https://doi.org/10.1186/1471-2164-13-374

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550–550. https://doi.org/10.1186/s13059-014-0550-8

Navarro, L., Zipfel, C., Rowland, O., Keller, I., Robatzek, S., Boller, T., & Jones, J. D. G. (2004). The transcriptional innate immune response to flg22. Interplay and overlap with Avr gene-dependent defense responses and bacterial pathogenesis. Plant Physiology, 135(2), 1113–1128. https://doi.org/10.1104/pp.103.036749

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols, 11(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology, 33(3), 290–295. https://doi.org/10.1038/nbt.3122

Ploetz, R. C. (1994). Panama disease: Return of the first banana menace. International Journal of Pest Management, 40(4), 326–336. https://doi.org/10.1080/09670879409371908

Ploetz, R. C. (2006). *Fusarium* wilt of banana is caused by several pathogens referred to as *Fusarium oxysporum* f. Sp. *cubense*. Phytopathology®, 96(6), 653–656. https://doi.org/10.1094/PHYTO-96-0653

Ploetz, R. C. (2015). *Fusarium* wilt of banana. Phytopathology®, 105(12), 1512–1521. https://doi.org/10.1094/PHYTO-04-15-0101-RVW

Sismak, S. B., & Zheng, S. (2018). Banana *fusarium* wilt (*Fusarium oxysporum* f. Sp. *cubense*) control and resistance, in the context of developing wilt-resistant bananas within sustainable production systems. Horticultural Plant Journal, 4(5), 208–218. https://doi.org/10.1016/j.hpj.2018.08.001

Spies, D., Renz, P. F., Beyer, T. A., & Ciaudo, C. (2019). Comparative analysis of differential gene expression tools for RNA sequencing time course data. Briefings in Bioinformatics, 20(1), 288–298. https://doi.org/10.1093/bib/bbx115

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), 9440–9440. https://doi.org/10.1073/pnas.1530509100

Sun, J., Zhang, J., Fang, H., Peng, L., Wei, S., Li, C., Zheng, S., & Lu, J. (2019). Comparative transcriptome analysis reveals resistance-related genes and pathways in *Musa acuminata* banana "Guijiao 9" in response to *Fusarium* wilt. Plant Physiology and Biochemistry, 141, 83–94. https://doi.org/10.1016/j.plaphy.2019.05.022

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLOS ONE, 6(7), e21800.

Swarupa, V., Ravishankar, K. V., & Rekha, A. (2014). Plant defense response against *Fusarium oxysporum* and strategies to develop tolerant genotypes in banana. Planta, 239(4), 735–751. https://doi.org/10.1007/s00425-013-2024-8

Tsuda, K., & Katagiri, F. (2010). Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. Current Opinion in Plant Biology, 13(4), 459–465. https://doi.org/10.1016/j.pbi.2010.04.006

Wang, Z., Zhang, J., Jia, C., Liu, J., Li, Y., Yin, X., Xu, B., & Jin, Z. (2012). De Novo characterization of the banana root transcriptome and analysis of gene expression under *Fusarium oxysporum* f. Sp. *cubense* tropical race 4 infection. BMC Genomics, 13(1), 650–650. https://doi.org/10.1186/1471-2164-13-650

Zhang, L., Yuan, T., Wang, Y., Zhang, D., Bai, T., Xu, S., Wang, Y., Tang, W., & Zheng, S.-J. (2018). Identification and evaluation of resistance to *Fusarium oxysporum* f. Sp. *Cubense* tropical race 4 in *Musa acuminata* Pahang. Euphytica, 214(7), 106–106. https://doi.org/10.1007/s10681-018-2185-4

Zuo, C., Deng, G., Li, B., Huo, H., Li, C., Hu, C., Kuang, R., Yang, Q., Dong, T., Sheng, O., & Yi, G. (2018). Germplasm screening of *Musa* spp. For resistance to *Fusarium oxysporum* f. Sp. *Cubense* tropical race 4 (Foc TR4). European Journal of Plant Pathology, 151(3), 723–734. https://doi.org/10.1007/s10658-017-1406-3

APPENDIX A SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Appendix A.1 - Supplementary figures

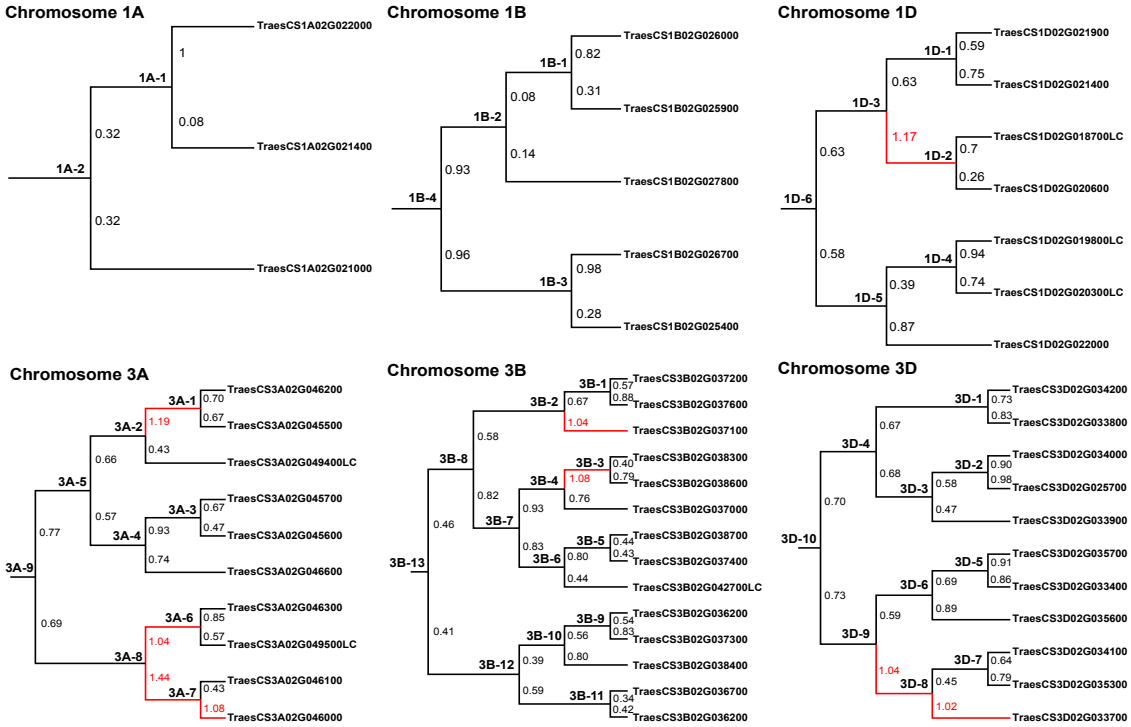


Figure S2.1 Ka/Ks phylogenetic tree of common wheat BBIs separated by chromosome. The values on each branch indicate the ratio for that pair of genes. Branches and values greater than one are highlighted in red.

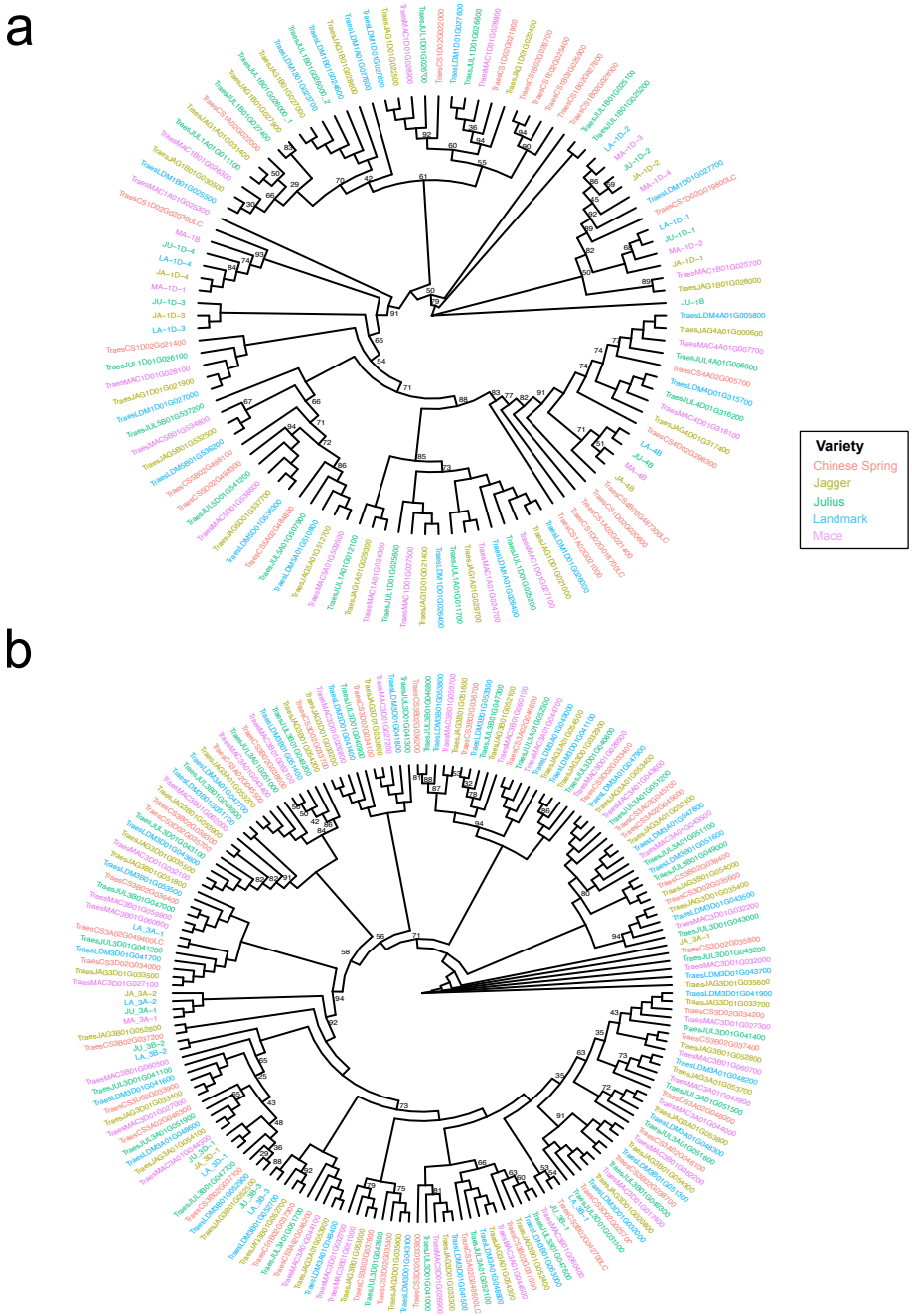


Figure S2.2 Phylogenetic tree of BBIs identified in common wheat landrace 'Chinese Spring' and four common wheat varieties ('Jagger', 'Mace', 'Landmark' and 'Julius') on **a** homoeologous group 1, 4, and 5 chromosomes and **b** homoeologous group 3 chromosomes**.** The trees were built with the model (WAG+G4) which has the lowest BIC value using 1000 bootstrap replications. Only bootstrap support values below 95 are indicated on the tree. Genes are color-coded based on wheat variety.

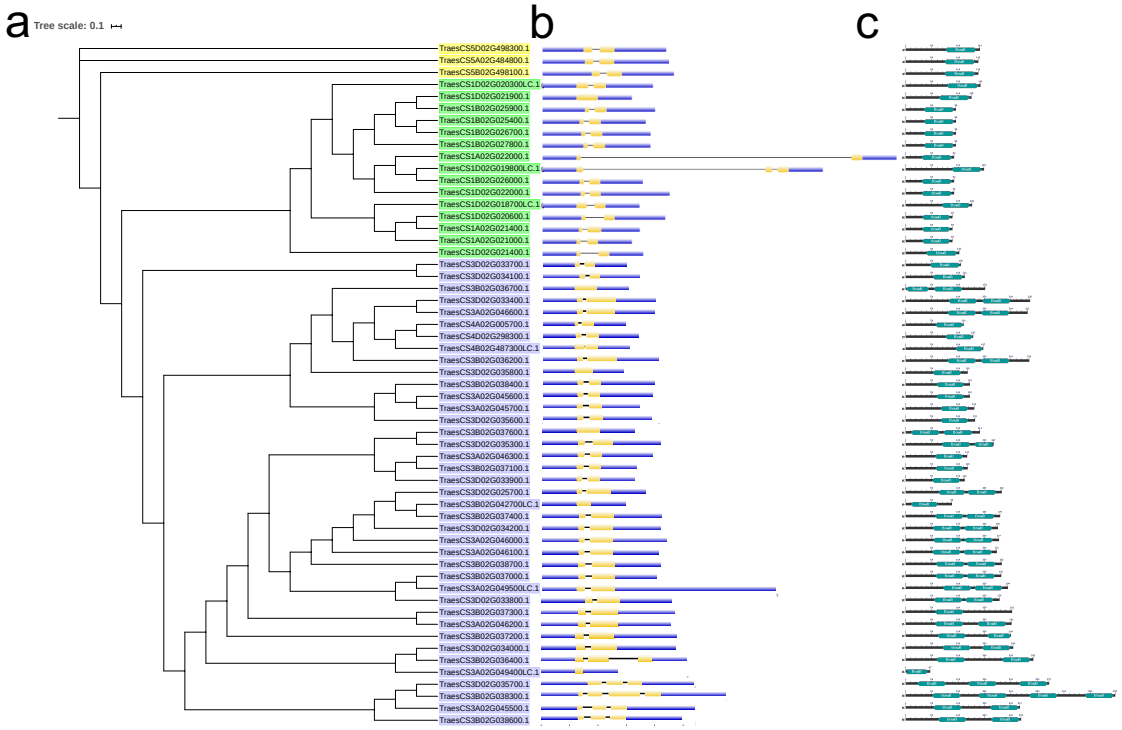


Figure S2.3 Structural characterization of common wheat BBIs. **a** Phylogenetic tree of 57 wheat BBI genomic sequences. The alignment was conducted with IQ-TREE to predict best fit model for nucleic acid with the lowest BIC value. Gene names are color coded to indicate different clades which were grouped based on their nucleic acid structure. **b** Intron-exon structure of each BBI gene predicted by comparison of CDS and gDNA sequence. Blue rectangles indicate untranslated regions, black lines indicate introns and yellow rectangles indicate exons. **c** Functional domain discovery, the Bowman-Birk domain prediction was conducted by NCBI-CDD to look for smart00269 (Bowman-Birk type protease inhibitor from SMART database). Blue rectangles indicate BBI functional domains and black lines indicate other amino acids.

Appendix A.2 – Supplementary tables for chapter 2 (.xls)

Table S2.1 List of 62 common wheat BBIs in the IWGSC RefSeq v1.1 genome assembly, and five additional genes that were excluded due to the lack of a complete BBI domain. Information includes their gene position (Gene ID based on IWGSC RefSeq v1.1 gene models, name based on their homoeologous relationships, chromosome locations and order), gene structure and features (number of exons and BBI domains, BBI domain evolutionary model types (Mello et al., 2003), amino acids at P1-P1` motif position, number of complete BBI domains with all required Cys residues, protein length and molecular weight), signal peptide prediction (SP prediction as signal peptide or other, prediction confidence, predicted cleavage site position, and + = present, – = absent), pseudogene prediction (T = True, F = False), and log_2TPM values of expression during development and log_2 Fold-change TPM of biotic and abiotic stress expression datasets.

Table S2.2 List of BBI genes in *T. aestivum, Ae. tauschii, T. urartu, T. dicoccoides*, and *H. vulgare* for which manual curation was performed. Details of the position and confidence level of the signal peptide site are included for both the original predicted sequence and the manually curated sequence. Full details of the manual curation are provided in column K, which have been corrected for the initiation codon. All curated sequences have a signal peptide prediction greater than 0.97. BBI genes with abnormal N-terminal truncation were also listed in column G.

Table S2.3 List of six putative wheat BBIs identified in previous studies. Information includes their original and alternative gene names, corresponding protein ID in the UniProt database, e-value for HMMscan of the BBI domain, protein sequences documented in the UniPort database, complete protein sequences based on their annotation in the IWGSC RefSeq v1.1 genome assembly and citations for the studies where these proteins were originally reported.

Table S2.4 Common wheat BBI homoeologous groups divided by chromosome.

Table S2.5 List of BBIs identified from *O. sativa, Z. mays, B. distachyon, H. vulgare, Ae. tauschii, T. urartu* and *T. dicoccoides.* Information includes species, gene number named by order of the gene ID from the source model, alternative name and the citation for where the name was first described, gene ID, chromosome position, BBI domain type, protein length, number of BBI domains, source of genome assembly and gene ID converter from IRGSP-1.0 to MSU for rice BBIs. For BBIs without alternative names and with uncharacterized model types, we used '-' symbol.

Table S2.6 Homologous relationships of BBIs in common wheat compared to *T. urartu* (A genome), *Ae. tauschii* (D genome) and *T. dicoccoides* (AB genomes). Orthologous genes are presented in the same row. BBIs with uncharacterized homologous relationships were placed in separate rows and labelled "ungrouped".

Table S2.7 List of BBIs identified in common wheat cultivars 'Jagger', 'Landmark', 'Julius' and 'Mace'. Information includes their gene ID according to the 10+ wheat genome project annotation (Walkowiak et al., 2020), chromosome and positions and their projections in 'Chinese Spring' where available. BBI genes present in some cultivars but absent from 'Chinese Spring' were named based on the cultivar (e.g. JA means 'Jagger', JU means 'Julius') followed by their chromosome and the physical order on that chromosome based on our *de novo* ORF prediction. For example, *JA_1D-1* refers to the first BBI gene on 'Jagger' chromosome 1D that is absent from the 'Chinese Spring' reference assembly.

Table S2.8 List of two common wheat BBIs identified in the "Triticum 4.0" assembly of 'Chinese Spring', but absent from the IWGSC RefSeq v1.1 assembly. Information includes their gene name, chromosomal location, corresponding orthologous gene from the IWGSC RefSeq v1.1 assembly, gene structure and features (exon and domain numbers, model types, P1-P1` motif residues, protein length and molecular weight) and signal peptide prediction.

Table S2.9 Functional annotation and genomic position of all genes 200 kb upstream and downstream of BBI clusters on homoeologous group 1 and 3 chromosomes. Information includes gene ID for both high and low confidence genes, their location on each chromosome, and their functional annotation and Pfam domains based on IWGSC RefSeq v1.1 gene models (Appels et al., 2018). BBI genes identified in our study are highlighted in red and genes annotated as other trypsin inhibitors are highlighted in blue.

Table S2.10 Number of genes sharing functional annotation terms from IWGSC RefSeq v1.1 gene models 200 kb upstream and downstream of BBI clusters on homoeologous group 1 and 3 chromosomes. The number of BBIs on each chromosome is highlighted in red. Gene number is based on descriptive annotations from gene models.

Because some BBI genes identified in our study are annotated as ‘trypsin inhibitor’ in these gene models, there is a slight discrepancy between the number of BBI genes described in this table and the total number of BBIs.

APPENDIX B SUPPLEMENTARY MATERIALS FOR CHAPTER 3

Appendix B.1 - Supplementary figures

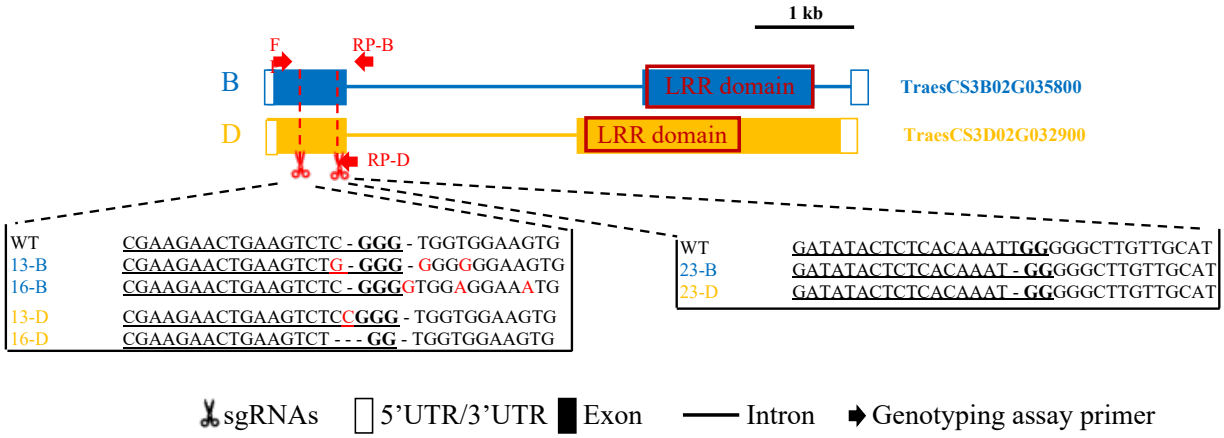


Figure S3.1. Functional validation of *TraesCS3B02G035800* (annotated as *RPM1*) in wheat. Two sgRNAs, marked as scissors, were designed targeting *TraesCS3B02G035800* and its homoeologous copy *TraesCS3D02G032900* to generate gene knockout mutants with CRISPR/Cas 9. The Leucine-rich repeat (LRR) domains were marked with red box. Genotyping assay primers included a common forward primer (FP) and genome specific reverse primers (RP-B and RP-B), and they were marked as arrows

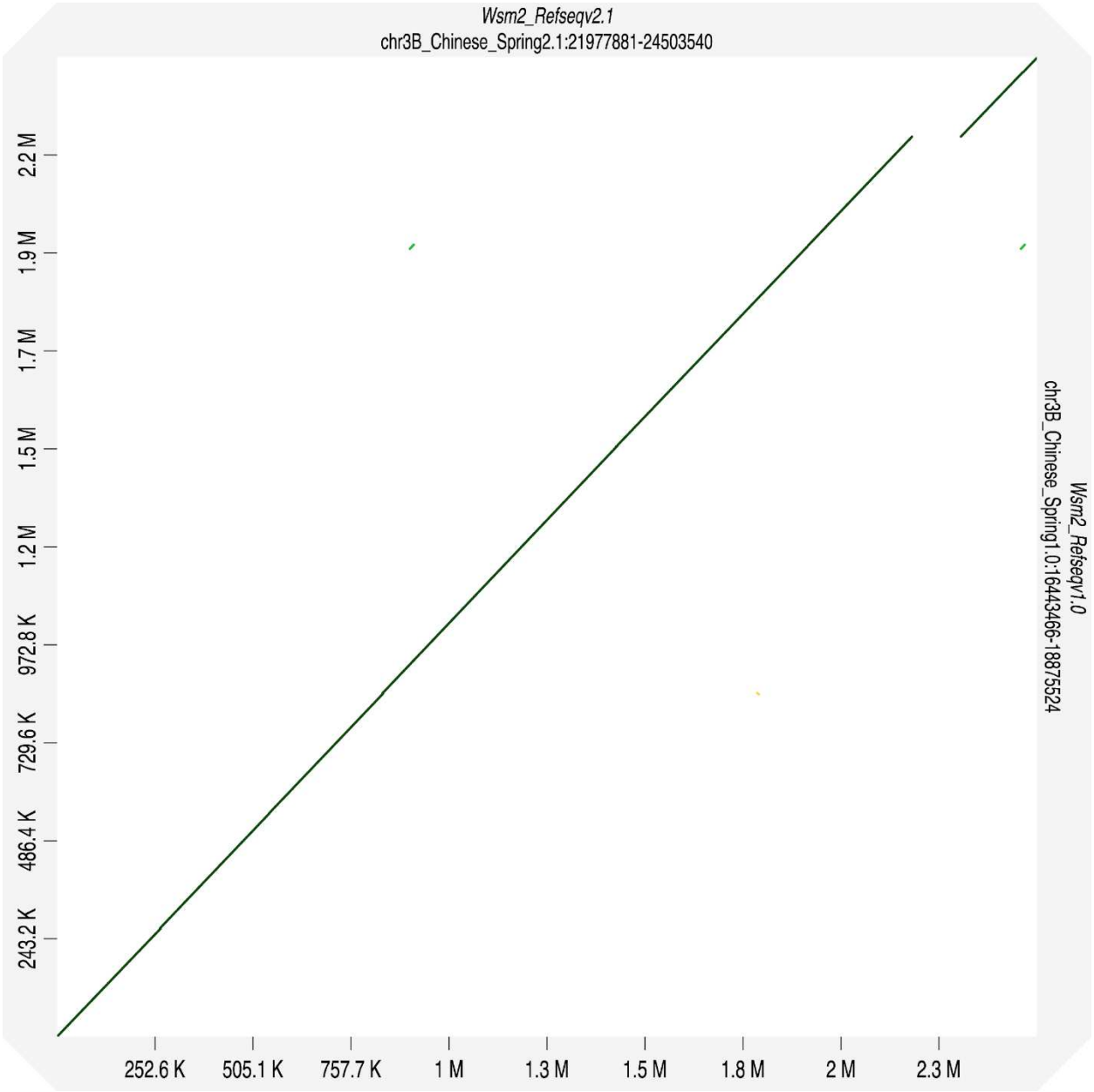


Figure S3.2. Dot plot for *Wsm2* locus in RefSeq v1.0 versus RefSeq v2.1. X-axis represented RefSeq v2.1, Y-axis indicate RefSeq v1.0. The dot plot was based on genomic DNA sequence of *Wsm2*. Green line indicated identify 0.95-1 percentage. The graph was made with D-genies.

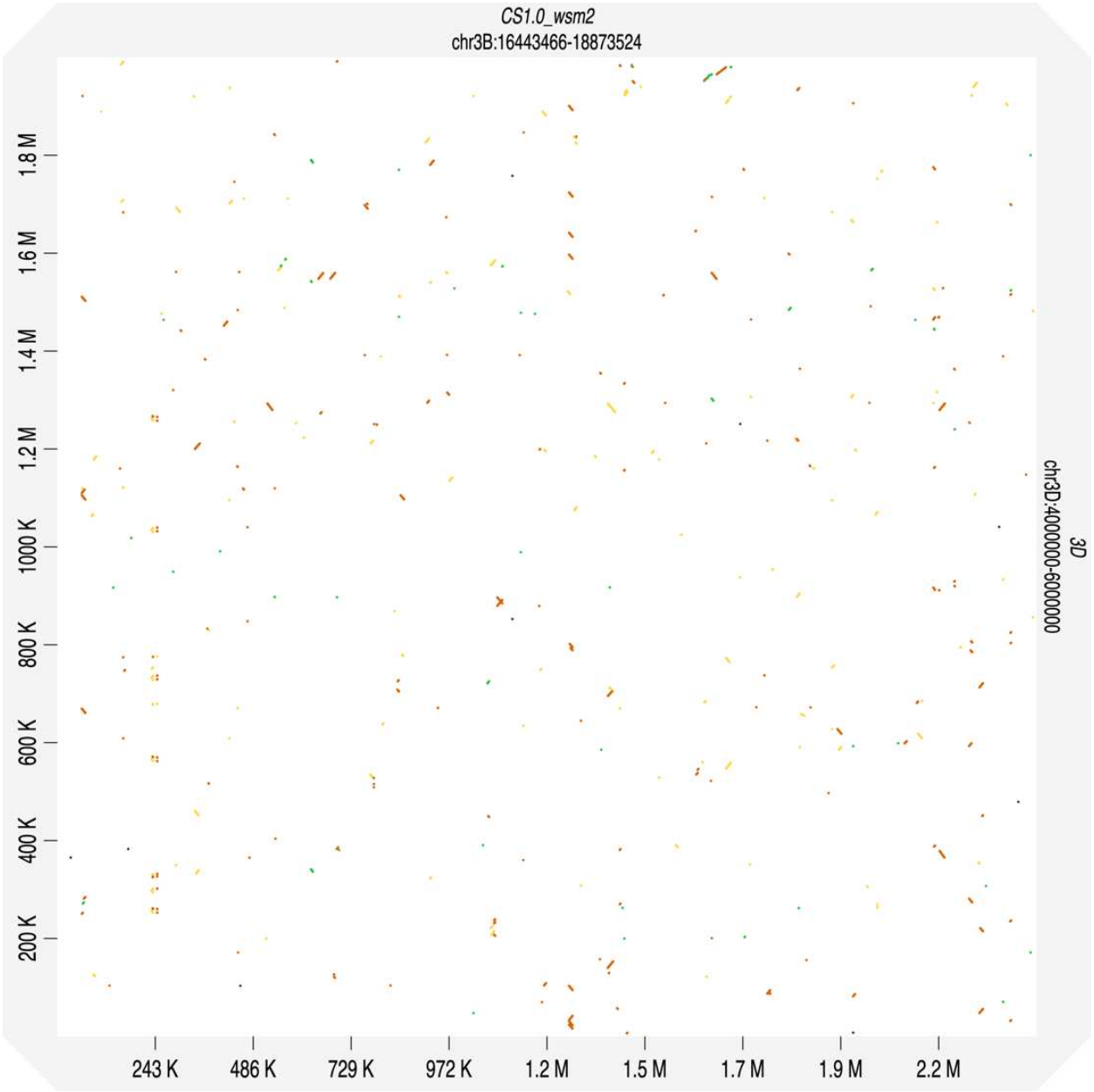


Figure S3.3. Dot plot for genomic sequence between significant QTL markers on 3D (4 Mb – 6 Mb) and the 2.4 Mb *Wsm2* on 3B.

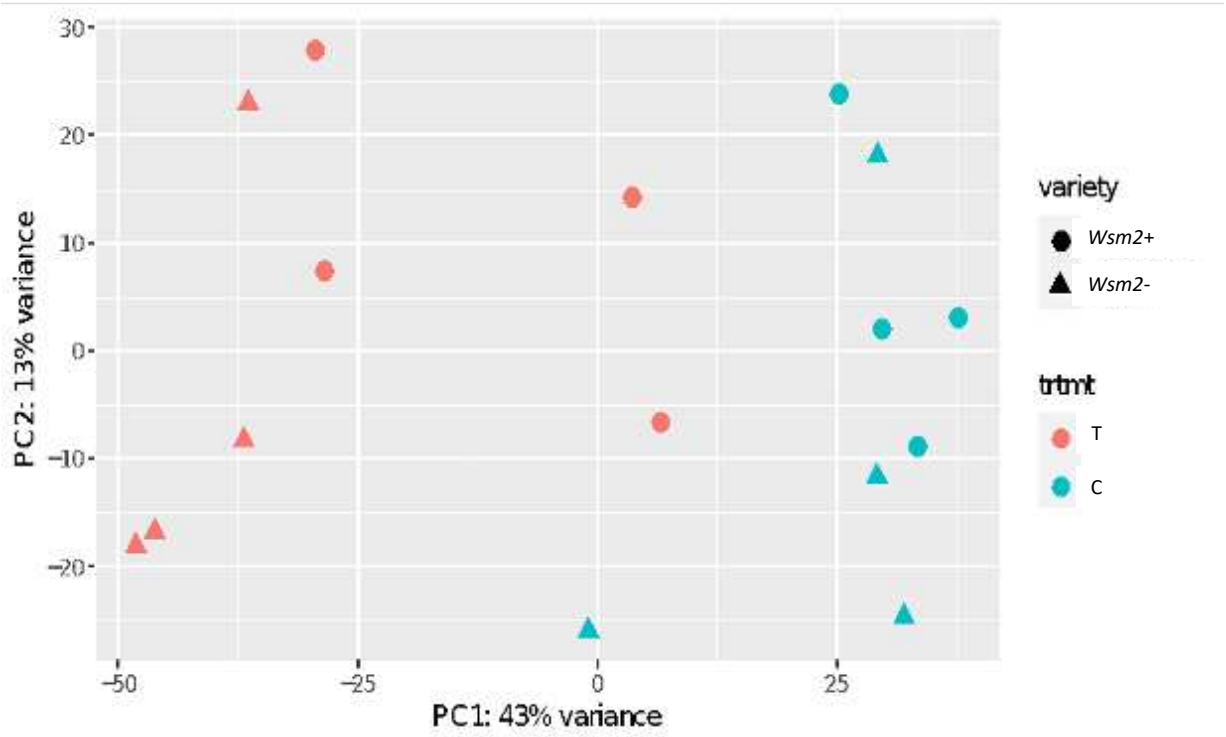


Figure S3.4. PCA plots for 16 samples in the RNA-seq experiments**.** Variety types are indicated by shape, circle means *Wsm2+* whereas triangle means *Wsm2-*; treatment type are indicated by color, red means WSMV inoculation treatment (T) whereas blue means mock inoculation treatment (C). Each variety and treatment combination have four biological replicates.

Appendix B.2 – Supplementary tables for chapter 3 (.xls)

Table S3.1. Haplotype and phenotype for selected doubled haploid individuals used for the RNA-seq study.

Table S3.2. Information for four KASP markers from Tan et al., 2017 paper.

Table S3.3. Primer and probe information for WSMV amplification, qRT-PCR gene expression, validating CRISPR/Cas 9 construct and genotyping RPM1 edits.

Table S3.4. Homoeologues for TraesCS3B02G035800 and primers used to synthesis sgRNAs.

Table S3.5. Physical position of four *Wsm2* associated KASP markers in the wheat pangenomes.

Table S3.6. Information for 94 candidate genes within *Wsm2*, including their physical position mapped to IWGSC RefSeq v1.0 and IWGSC RefSeq v2.1, gene ID and protein annotation.

Table S3.7. Significant GBS markers (LOD > 3) from the linkage mapping analysis.

Table S3.8. Polymorphisms and genetic variations underlying *Wsm2* locus in 'Snowmass' when mapped to IWGSC RefSeq v1.0.

Table S3.9. Read and mapping statistics of RNAseq samples using IWGSC RefSeq v1.0 or it's combination with WSMV genome as reference.

Table S3.10. Information for 16 RNA seq samples and the counts per million (CPM) value for the presence of WSMV genome in each sample.

Table S3.11. Information for DEGs between treatment and between genotype using IWGSC RefSeq v1.0 as reference.

Table S3.12. GO enrichment and significant GO terms ($p < 0.01$) for up- and down-regulated DEGs between WMSV treated versus mock treated samples and for the 3470 unique DEGs between genotype at WSMV treated conditions.

Table S3.13. Position and annotation for the 22 DEGs on chromosome 3B in the C_{Wsm2-}^{Wsm2+} comparisons. DEGs near *Wsm2* markers are highlight with red.

Table S3.14. Information for DEGs from pairwise comparisons when samples were mapped to *de novo* assembly of unmapped transcriptomes.

Table S3.15. Information for differentially expressed genes from unmapped transcriptomes that overlapped between T_{Wsm2-}^{Wsm2+} and $Wsm2\ +_{C}^{T}$.

Table S3.16. Presence and absence analysis against wheat pangenomes for three unique transcripts identified from *de novo* assembly of unmapped transcriptomes.

Table S3.17. Gene expression (TPM, log2 fold change, and padj) for the 94 candidate genes underlying *Wsm2*.

Table S3.18. Blastn output for unique transcripts from the *de novo* transcriptome assembly of unmapped reads in the *Wsm2*+ samples against pangenome wheat cultivars.

Table S3.19. Genotyping and edit effects of CRISPR/Cas 9 edited wheat materials.

APPENDIX C SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Appendix C.1 - The R script used to run WGCNA analysis (wgcna_standard.R).

Appendix C.2 – Supplementary tables for chapter 4 (.xls)

Table S4.1. Summary of the 24 RNA-seq samples**.** Libraries were named AC1 to AC24, and details include line ID, line name, time point (Day 0, 1, 3, or 7 post inoculation) and resistance or susceptibility to *Foc*-STR4.

Table S4.2. Summary statistics of the RNA seq samples reads and alignment rates. The results for overall mapping rate, unique mapping rate, multi-mapping rate, and unmapped reads comparing two alignment tools, STAR and HISAT2, were included.

Table S4.3. Modules identified from WGCNA analysis. The number of DEGs in each module, eigengene expression profile at each time point, the top enriched GO terms associated with MF and BP for each module, and hub gene information**.**

Table S4.4. Significantly enriched GO terms ($P < 0.01$) associated with BP and MF for DEGs in each module identified from the WGCNA analysis.

Table S4.5. Expression values and DEG statistics (log_2 fold change with and without shrinkage and *P* adj) for DEGs between resistant versus susceptible samples within each time point at T0, T1, T3 and T7.

Table S4.6. Significantly enriched GO terms ($P < 0.05$) associated with BP, MF, and CC for up- and down-regulated DEGs between genotype at each time point.

Table S4.7. Expression values of 59 candidate genes from the gene family identified underlying the novel QTL. The expression values were shown in transcript per million (TPM) for both resistant and susceptible genotype at all four time points.