Genetic dissection of maize regeneration and wheat disease resistance

by

Guifang Lin

B.S., Fujian Agriculture and Forestry University, China, 2011
M.S., Fujian Agriculture and Forestry University, China, 2014

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

The growing human population worldwide and the changing growth environments require significant crop improvement, which can be accelerated by plant genome engineering. Developing plant cultivars amenable to transformation and improving understanding of the genetic bases of important phenotypic traits can facilitate the use of advanced genome engineering technologies. This dissertation is focused on the genetic analysis of maize transformation and wheat resistance to the disease of leaf rust. The results will provide knowledge to improve crop transformation and wheat disease resistance.

Plant transformation is a powerful tool for crop improvement and gene function validation. However, the transformation efficiency of maize is highly dependent on the tissue types and the genotypes. The maize inbred line A188 is amenable to transformation. A188 also exhibits many contrasting traits to the inbred line B73, which is recalcitrant to transformation. B73 was used to generate the first maize reference genome. The lack of genome sequences of A188 limits the use of A188 as a model for functional studies. Here, a chromosome-level genome assembly of A188 was constructed using long reads and optical physical maps. Genome comparison of A188 with B73 based on both whole genome alignments and sequencing read depths identified approximately 1.1 Gb syntenic sequences as well as extensive structural variation. Further, transcriptome and epigenome analyses with the A188 reference genome revealed enhanced gene expression of defense pathways and altered DNA methylation patterns of embryonic callus. The A188 genome assembly provides a foundational resource for analyses of genome variation and gene function in maize. In maize, morphologic types of calli induced from immature embryos are associated with the regeneration capability, which is a major factor determining the transformation efficiency. Here, two contrasting callus types, slow-growth type I calli and fast-growth type II calli, from the

selected B73xA188 F2 population were sequenced using Genotyping-By-Sequencing (GBS) and RNA-Seq. With both approaches, the genomic loci associated with the callus type were mapped to chromosomes 2, 5, 6, 8, and 9. From F2 RNA-Seq, differentially expressed genes were identified from the comparison of type II and I calli. In addition, RNA-Seq analysis was performed using fast- and slow-growth calli identified for the A188 calli. Gene ontology (GO) enrichment analysis showed that the down-regulated genes in type II F2 calli and fast-growth A188 calli, as respectively compared to type I calli and slow-growth A188 calli, are overrepresented in the pathway related to cell wall organization, suggesting the role of cell wall formation in the callus development.

Besides maize genetic and genomic studies, the dissertation includes the cloning of a leaf rust resistance gene in wheat. Wheat leaf rust disease is caused by a fungal pathogen, *Puccinia triticina.* The *Lr42* gene from the wheat wild relative *Aegilops tauschii* confers resistance to all leaf rust races tested to date. Through bulked segregant RNA-Seq (BSR-Seq) mapping and further fine mapping, we identified an *Lr42* candidate gene, which encodes a nucleotide-binding site leucine-rich repeat (NLR) protein. Transformation of the candidate gene to a leaf rust-susceptible wheat cultivar markedly enhanced the disease resistance, confirming the candidate NLR gene is the *Lr42* gene. Cloning of *Lr42* expands the repertoire of cloned rust resistance genes, as well as provides precise diagnostic gene markers for wheat improvement.

Genetic dissection of maize regeneration and wheat disease resistance

by

Guifang Lin

B.S., Fujian Agriculture and Forestry University, China, 2011
M.S., Fujian Agriculture and Forestry University, China, 2014

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Dr. Sanzhen Liu

Copyright

Abstract

The growing human population worldwide and the changing growth environments require significant crop improvement, which can be accelerated by plant genome engineering. Developing plant cultivars amenable to transformation and improving understanding of the genetic bases of important phenotypic traits can facilitate the use of advanced genome engineering technologies. This dissertation is focused on the genetic analysis of maize transformation and wheat resistance to the disease of leaf rust. The results will provide knowledge to improve crop transformation and wheat disease resistance.

Plant transformation is a powerful tool for crop improvement and gene function validation. However, the transformation efficiency of maize is highly dependent on the tissue types and the genotypes. The maize inbred line A188 is amenable to transformation. A188 also exhibits many contrasting traits to the inbred line B73, which is recalcitrant to transformation. B73 was used to generate the first maize reference genome. The lack of genome sequences of A188 limits the use of A188 as a model for functional studies. Here, a chromosome-level genome assembly of A188 was constructed using long reads and optical physical maps. Genome comparison of A188 with B73 based on both whole genome alignments and sequencing read depths identified approximately 1.1 Gb syntenic sequences as well as extensive structural variation. Further, transcriptome and epigenome analyses with the A188 reference genome revealed enhanced gene expression of defense pathways and altered DNA methylation patterns of embryonic callus. The A188 genome assembly provides a foundational resource for analyses of genome variation and gene function in maize. In maize, morphologic types of calli induced from immature embryos are associated with the regeneration capability, which is a major factor determining the transformation efficiency. Here, two contrasting callus types, slow-growth type I calli and fast-growth type II calli, from the

selected B73xA188 F2 population were sequenced using Genotyping-By-Sequencing (GBS) and RNA-Seq. With both approaches, the genomic loci associated with the callus type were mapped to chromosomes 2, 5, 6, 8, and 9. From F2 RNA-Seq, differentially expressed genes were identified from the comparison of type II and I calli. In addition, RNA-Seq analysis was performed using fast- and slow-growth calli identified for the A188 calli. Gene ontology (GO) enrichment analysis showed that the down-regulated genes in type II F2 calli and fast-growth A188 calli, as respectively compared to type I calli and slow-growth A188 calli, are overrepresented in the pathway related to cell wall organization, suggesting the role of cell wall formation in the callus development.

Besides maize genetic and genomic studies, the dissertation includes the cloning of a leaf rust resistance gene in wheat. Wheat leaf rust disease is caused by a fungal pathogen, *Puccinia triticina.* The *Lr42* gene from the wheat wild relative *Aegilops tauschii* confers resistance to all leaf rust races tested to date. Through bulked segregant RNA-Seq (BSR-Seq) mapping and further fine mapping, we identified an *Lr42* candidate gene, which encodes a nucleotide-binding site leucine-rich repeat (NLR) protein. Transformation of the candidate gene to a leaf rust-susceptible wheat cultivar markedly enhanced the disease resistance, confirming the candidate NLR gene is the *Lr42* gene. Cloning of *Lr42* expands the repertoire of cloned rust resistance genes, as well as provides precise diagnostic gene markers for wheat improvement.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest appreciation to many people who supported me to finish this dissertation. First, I would like to express my thanks to my mentor, Dr. Sanzhen Liu, for giving me the great opportunity to work on many fantastic projects. Thank him so much for all the encouragements! He always has the magic to grab me out when I get stuck in research. Then, I would like to express my sincere gratitude to the dissertation committee, Drs. Christopher Toomajian, Guihua Bai, Jesse Poland, Robert Bowden, and the chair, Thomas Platt for their guidance on the research and the help on the writing. Thank them so much for the time!

I would like to acknowledge all my colleagues in the Liu Lab. Thanks to them for the field work to build maize population, the data analysis, and all the other help to finish the research work. In addition, I would like to express my appreciation to many groups that contributed to the projects. Thanks to the groups of Drs. Sunghun Park, Frank White, Hairong Wei, Myeong-Je Cho in the NSF PREG team and Dr. Guoying Wang at Chinese Academy of Agricultural Sciences for the help to finish the maize regeneration projects. Thanks to the groups of Drs. Bikram Gill, Guihua Bai, Harold N. Trick, Robert L. Bowden, and Jesse Poland at Kansas State University for the work to finish the wheat leaf rust gene cloning project. Also, I would like to thank Dr. Barbara Valent group for the help on the wheat blast project, which was not presented in the dissertation. Without their support, I was unable to finish the work.

I would like to express my special thanks to the Department of Plant Pathology at k-state. Thanks to the professors for all the professional training, thanks to all the office staff for their help, and thanks to all the people who gave me support and big smiles around the building!

Finally, I would like to thank my family for their full support! Special thanks to all my friends who help me to take care of my little boy during the pandemic! Thanks to all!

Chapter 1 - General Introduction

Abstract

The global yield of maize and wheat urgently needs to be increased to face the challenges of growing population and climate changes. Breeding cultivars adapting to diverse environments is one of the keys to increase the crop production. Maize is a monoecious grass species and was domesticated from teosinte. In addition to being the top cereal crop in the US, maize is also an attractive model organism for plant research. Genetic and genomic resources constructed and utilized for research studies and maize improvement are introduced in this chapter. Bread wheat is the major wheat crop and an allohexaploid grass species, originated from the hybridization of emmer and goatgrass. Emmer and goatgrass still exist in the wild, which can serve as the genetic pool for wheat breeding. Yield loss caused by plant pathogens is one of the constraints of wheat productivity, and breeding genetic resistant cultivars can be an economical and efficient strategy to reduce the loss. This chapter briefly introduces wheat domestication and the strategy to utilize wild wheat ancestors to improve disease resistance durability. More than 35 maize and more than 19 wheat (including its close relatives) genome assemblies have been generated even though both genomes are large and contain more than 80% repetitive sequences. These great achievements are attributable to the revolutionary development of DNA sequencing technologies. As the sequencing technology continues to improve, more high-quality crop genomes will be produced to facilitate genomic analysis and crop improvement. This chapter briefs the current development of sequencing technologies.

Introduction

Maize and wheat are the two most important cereal crops feeding billions of people in the world (García-Lara and Serna-Saldivar 2019). Facing the increasing population and climate change, the production of maize and wheat needs to be increased with modern breeding technologies. In modern breeding, plant breeders and researchers are combining diverse genetic resources and genomic technologies to speed up the breeding cycles to produce high yield varieties that can be adapted to diverse environment conditions (Tester and Langridge 2010; Watson *et al.* 2018; Ahmar *et al.* 2020; Gosal and Wani 2020). In this chapter, the general information of the maize and wheat and the latest sequencing technologies for genomic research are briefly reviewed.

Maize is both an important crop a model organism for basic research

Maize is a grass species native to America

Maize (*Zea mays* L. ssp. *mays*), also known as corn, is widely utilized as a staple crop, an important source for animal feed, and a bioethanol resource (Shiferaw *et al.* 2011). The high consumption demand and the high calorie production of maize make it one of the most agricultural important cereal crops worldwide. Monoecious maize, a member of the grass family Poaceae, is believed to domesticate from its wild ancestor, teosinte, about 9,000 years ago in Mexico (Benz 2001; Matsuoka *et al.* 2002; Doebley *et al.* 2006). During the domestication, some genes, particularly transcription factor genes, such as *tb1* (*Teosinte branched 1*) and *tga1* (*Teosinte glume architecture 1*), played key roles in transforming teosinte to the maize crop (Doebley *et al.* 1997; Wang *et al.* 2005; Liu *et al.* 2020). After domestication in central America, maize was spread to the world along with human activities (Prasanna 2012). The yield of maize was dramatically increased after the adoption of hybrids in the 20th century (Crow 1998; Duvick 2005). Until now,

the hybrids are still dominating in the corn industry. Transforming the wild teosinte to the high yield hybrid maize shows the great potential of breeding improvement to solve the problem of food security.

Maize serves as an important model organism

Besides the agricultural value, maize is an important model organism for basic plant research. Even though maize has a longer growth period and larger plant size than many other plant model organisms (Nannas and Dawe 2015), there are many advantages for maize to be an important model organism. For instance, the separate male and female flowers in a single individual make it easy to outcross but also preserves the ability of self-pollination; the strong hybrid vigor makes it an attractive model to understand the genetics of heterosis; the large scale of chromosomes allows effective cytogenetic analysis; the large kernel arranged together facilitates genetic studies of kernel development; and the availability of rich genetic and genomic resources enables the genetic dissection of quantitative traits (Strable and Scanlon 2009; Nannas and Dawe 2015; Hake and Ross-Ibarra 2015).

Genetic resources of maize

Germplasms are the fundamental resources for both breeding and research. Genetic analysis of the US commercial lines from 1980 to 2008 showed the commercial germplasms originated from both public and private breeding programs (Mikel and Dudley 2006; Mikel 2011). For quantitative analysis of maize, the Nested Association Mapping population (NAM) was released for the maize community along with individual genotype data (McMullen *et al.* 2009). In the NAM population, the agriculturally important and well-studied reference line B73 was crossed with 25 diverse inbred lines to produce 200 recombinant inbred lines (RILs) per family, resulting

in 5000 RILs for the whole population. Until 2020, the NAM population has been used for about 22 publications studying more than 100 traits (Gage *et al.* 2020). The impact of the NAM design is not limited to maize genetic studies. In addition to the NAM population, a maize association population with 282 inbred lines and Ames inbred lines panel with 2,815 accessions are also available for the public access along with genomic resources (Flint-Garcia *et al.* 2005; Romay *et al.* 2013). Other than quantitative analysis, the UniformMu population of mutants with insertions of the *Mutator* transposon in W22 background provides genetic materials for functional analysis (McCarty *et al.* 2005).

Genomic resource of maize

The reference genome of the elite inbred line B73 was first assembled with bacterial artificial chromosomes (BACs) sequencing (Schnable *et al.* 2009) and intensely used for maize genetics and functional studies (Chia *et al.* 2012; Hufford *et al.* 2012; Jiao *et al.* 2012; Romay *et al.* 2013; Hirsch *et al.* 2014; Li *et al.* 2014, 2015; Mao *et al.* 2015). With the rapid development of sequencing technologies, the quality of the B73 reference genome is continually being improved (Jiao *et al.* 2017; Hufford *et al.* 2021). The version 4 of the B73 reference genome consists of 2.1 Gb and 39,590 genes. The reference genome facilitates the discovery of structural variation among cultivars within the maize species (Springer *et al.* 2009; Hirsch *et al.* 2016; Sun *et al.* 2018), which inspires the research community to construct pan-genome references (Hirsch *et al.* 2014; Della Coletta *et al.* 2021). High-quality reference genomes from diverse lines can be the first step to build the pan-genome (Jayakodi *et al.* 2021). Recently, the complete genome assemblies of 25 founder lines of the maize NAM population were released (Hufford *et al.* 2021). Based on all the 25 newly released reference genomes and the updated B73 reference genome, 103,538 pan-genes

were identified(Hufford *et al.* 2021). Of the 103,538 pan-genes, about 58% are core genes and 32% are dispensable genes (Hufford *et al.* 2021). Besides NAM founders, many other important lines have been sequenced. The inbred lines PH207 and Mo17 were two of the major progenitors of the maize commercial hybrid from 1980-2004 (Mikel and Dudley 2006), and the genomes were assembled to 2.1 Gb and 2.2 Gb, respectively (Hirsch *et al.* 2016; Sun *et al.* 2018). The background line of UniformMu seed stock, W22, was sequenced and would be important for the UniformMu related function and transposon study (Springer *et al.* 2018). Genomes of four representative germplasms of European breeding programs, EP1, F7, DK105, and PE0075, were recently assembled (Haberer *et al.* 2020). The genome assembly of inbred line SK (small-kernel), which originated from a tropical landrace, can be used for yield trait analysis (Yang *et al.* 2019). The genome of inbred line Ia453 increased the number of sweet corn reference genomes (Hu *et al.* 2021). All these high-quality genome assemblies of the diverse germplasms will not only facilitate functional analysis of phenotypic traits, but also will be the foundation to build a high-coverage maize pan-genome.

Wheat and its disease resistance breeding

Wheat is a global food crop originating in the Fertile Crescent of the Middle East

Wheat is a daily calories and protein source for people all over the world (Shiferaw *et al.* 2013). The domestication and cultivation of wheat began about 10,000 years ago in Fertile Crescent (Bell 1987). At the beginning of wheat cultivation, the diploid domesticated einkorn (*Triticum monococcum*) and tetraploid domesticated emmer (*Triticum turgidum* ssp. *dicoccon*) were the two of the three staple cereals (Bell 1987). Eventually, an allohexaploid bread wheat (*Triticum aestivum* L.) arose from the hybridization of its ancestors (tetraploid *Triticum turgidum*

and diploid *Aegilops tauschii*) in the Fertile Crescent about 8,000 years ago (McFadden and Sears 1946; Dubcovsky and Dvorak 2007; Faris 2014), and bread wheat, consisting of three genomes A, B, and D, became the major wheat crop (Dubcovsky and Dvorak 2007). In history, the productivity of wheat was dramatically increased during the Green Revolution through the adoption of technologies and varieties, which saved many people in developing countries from hunger (Shiferaw *et al.* 2013). However, the increase of wheat yield is slow in recent years, and the global consumption demand of wheat is still growing (Shiferaw *et al.* 2013). To face the challenges, modern breeding based on the diverse germplasms and new technologies is urgently needed (Tester and Langridge 2010; "Technologies to boost breeding" 2018; Pont *et al.* 2019; Voss-Fels *et al.* 2019; Qaim 2020; Alotaibi *et al.* 2021).

The wild relatives for wheat disease resistance breeding

Wild relatives preserve the genetic diversity that was not captured during the origin and domestication crops, and the introgression of traits from wild relatives is important for the crop adaptation (Hajjar and Hodgkin 2007; Dempewolf *et al.* 2017; He *et al.* 2019; Khoury *et al.* 2020). Genetic diversity analysis suggested that the D genome of the bread wheat captured less diversity from its donor, *Ae. tauschii*, than the A and B genomes (Dubcovsky and Dvorak 2007). Therefore, there is a great potential to improve the agronomic traits of wheat with the genetic diversity of *Ae. tauschii* and its related species. However, the wild relatives' resource for modern wheat breeding is not only limited to the *Aegilops* genus (Wulff and Moscou 2014).

Disease caused by plant pathogens is one of the major constraints of wheat productivity (Shiferaw *et al.* 2013). The most efficient way to manage wheat disease is to grow the cultivars with genetic resistance or resistance genes (Bockus *et al.* 2001). To breed the resistanct cultivars,

it is crucial to identify the sources of genetic resistance genes. For wheat leaf rust, a fungal disease caused by *Puccinia triticina*, more than 79 resistance genes have been officially named (Park 2016; Bansal *et al.* 2017; Singla *et al.* 2017; Kolmer *et al.* 2018a; b; Qureshi *et al.* 2018; Kumar *et al.* 2021). Of the 79 *Lr* genes, 20 of them are from the wild relatives in the *Aegilops* genus, most of which have been transferred into *T. aestivum*, and 18 of them are from other wheat relatives of *T. aestivum* (Park 2016; Bansal *et al.* 2017; Singla *et al.* 2017; Kolmer *et al.* 2018a; b; Qureshi *et al.* 2018; Kumar *et al.* 2021). That a large number of *Lr* genes originated from wild wheat relatives supports the importance of introducing the diversity of wild relatives to enhance wheat disease resistance through breeding. Currently, more resistance genes are still needed to provide sufficient sources for breeding of the durable resistance.

Genomic resource of wheat and its wild relatives

Hexaploid wheat has a genome size of about 16 Gb consisting of three subgenomes, which is about 7 times that of the maize genome, and has 85% of repetitive sequences (Schnable *et al.* 2009; International Wheat Genome Sequencing Consortium (IWGSC) *et al.* 2018). Although the two progenitors of bread wheat, *Triticum turgidum* (AABB, ~12Gb) and *Ae. tauschii* (DD, ~4.3Gb) are tetraploid and diploid, respectively, their genome assemblies are still challenging (Citovsky and Keren 2017; Luo *et al.* 2017). With years of efforts and improved genomic technologies, the chromosome-scale genome assemblies of the bread wheat and its ancestors were currently available to public access (Citovsky and Keren 2017; Luo *et al.* 2017; International Wheat Genome Sequencing Consortium (IWGSC) *et al.* 2018). The genome of wild emmer (*T. turgidum ssp. dicoccoides*) was assembled into 10.1 Gb chromosomes and annotated with 65,012 high-confident gene models, while the genome of *Ae. tauschii* was assembled into 4.03 Gb

chromosomes and annotated with 39,622 high-confidence gene models (Citovsky and Keren 2017; Luo *et al.* 2017). The bread wheat was assembled to 14.5 Gb and annotated with 107,891 high-confidence gene models (International Wheat Genome Sequencing Consortium (IWGSC) *et al.* 2018). In addition to the bread wheat, the genome of durum wheat (*Triticum turgidum L. ssp. durum*) was assembled to 10.45 Gb (9.96 Gb chromosomes) and 66,559 high-confidence gene models (Maccaferri *et al.* 2019). Another milestone of the wheat genomic study is the release of 15 wheat genome assemblies, and structural variation was characterized among genomes (Walkowiak *et al.* 2020). The high quality of diverse genome assemblies will be a foundation resource for both pan-genome and molecular function analysis, which will eventually facilitate wheat breeding.

The long-read sequencing technologies facilitate the research of plant genomics

The long-read sequencing technology

The DNA sequencing technology is an invaluable invention for life science research, and it experienced revolutionary changes during the last several decades (Heather and Chain 2016; Shendure *et al.* 2017). The first generation sequencing represented by Sanger Sequencing contributed and was improved during the genome projects, such as the genome assembly projects of yeast, nematode, and human (Goffeau *et al.* 1996; C. elegans Sequencing Consortium 1998; International Human Genome Sequencing Consortium 2004; Shendure *et al.* 2017). The second generation sequencing, also named as next generation sequencing (NGS), is dominated by the Illumina platform, and the Illumina sequencing technology has more than 99.9% accuracy, the advantage of a high throughput, but with relative short sequencing reads (Shendure *et al.* 2017).

The emergence of long-read sequencing technologies was designed to overcome the limitations of read length of NGS (Heather and Chain 2016; Shendure *et al.* 2017; van Dijk *et al.* 2018).

Oxford Nanopore sequencing and PacBio SMRT sequencing are the two major long-read sequencing technologies and have been used to generate high quality maize and wheat genome assemblies (Jiao *et al.* 2017; Maccaferri *et al.* 2019; Haberer *et al.* 2020; Hufford *et al.* 2021; Hu *et al.* 2021). The basic principle of Nanopore sequencing is to trace the ion current changes in the protein pore when the single strand DNA goes through the pore (Jain *et al.* 2016), while the PacBio is to capture fluorescent pulses during the synthesis of a new complementary strand (Eid *et al.* 2009). The average read length of SMRT is 10-15 kb (van Dijk *et al.* 2018), while the mean length of Nanopore sequencing can be longer. Both technologies produce noisy reads that contain a relatively high error rate. Recently, the accuracy of PacBio reads has been improved to 99.8% by sequencing one molecule multiple times and generating the consensus sequence, producing high fidelity (HiFi) reads with very low errors (Wenger *et al.* 2019). Both technologies can sequence DNA base modifications (van Dijk *et al.* 2018), but Nanopore has the advantage to directly sequence RNA modifications (Parker *et al.* 2020). Overall, both technologies are being improved and are highly valuable for genomic research.

The application of long-read sequencing for genome assemblies

Plant genome assembly can be challenging due to the large genome size and highly repetitive features (Schnable *et al.* 2009; International Wheat Genome Sequencing Consortium (IWGSC) *et al.* 2018). Long reads, larger than 10kb, can overcome some limitations of short read performance in genome assemblies and variation discovery. In recent plant genome assemblies, the contig assembly with long reads becomes a common strategy (Song *et al.* 2020; Sun *et al.*

2020; Hufford *et al.* 2021; Hu *et al.* 2021). Although the error rate of long read sequencing is high, reads correction and contig polishing steps can improve the sequence accuracy (Jung *et al.* 2019; Amarasinghe *et al.* 2020). In addition to the important role for genome assembly, long reads can be used to discover structural variation associated with important agronomic traits (Chawla *et al.* 2021).

The application of long reads RNA-Seq

RNA sequencing plays a crucial role in transcriptome analysis (Stark *et al.* 2019). Long reads RNA-Seq technologies have been developed, including the Nanopore and PacBio RNA sequencing (Gonzalez-Garay 2016; Garalde *et al.* 2018). All these have been used for identification of transcripts, which provide evidence for gene discovery (Hufford *et al.* 2021), construction of transcriptome references (Minio *et al.*), and identification of long non-coding RNA (Cui *et al.* 2020). The Nanopore technology can sequence RNA directly, which has been utilized to characterize RNA modifications (Parker *et al.* 2020).

Conclusion

Crop improvement is crucial to feed the increasing population worldwide and relies on the genetic diversity of cultivars, land races and crop wild relatives. To fully utilize the resources, breeders and researchers need to efficiently identify, introduce and select genetic loci for desirable traits using advanced technologies. The development of sequencing technologies will be used to produce better genomic resources, which will accelerate the breeding.

References

Ahmar S., R. A. Gill, K.-H. Jung, A. Faheem, M. U. Qasim, *et al.*, 2020 Conventional and Molecular Techniques from Simple Breeding to Speed Breeding in Crop Plants: Recent Advances and Future Outlook. Int. J. Mol. Sci. 21. https://doi.org/10.3390/ijms21072590

Alotaibi F., S. Alharbi, M. Alotaibi, M. Al Mosallam, M. Motawei, *et al.*, 2021 Wheat omics: Classical breeding to new breeding technologies. Saudi J. Biol. Sci. 28: 1433–1444.

Amarasinghe S. L., S. Su, X. Dong, L. Zappia, M. E. Ritchie, *et al.*, 2020 Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21: 30.

Bansal M., S. Kaur, H. S. Dhaliwal, N. S. Bains, H. S. Bariana, *et al.*, 2017 Mapping of Aegilops umbellulata-derived leaf rust and stripe rust resistance loci in wheat. Plant Pathol. 66: 38–44.

Bell G. D. H., 1987 The history of wheat cultivation, pp. 31–49 in *Wheat Breeding: Its scientific basis*, edited by Lupton F. G. H. Springer Netherlands, Dordrecht.

Benz B. F., 2001 Archaeological evidence of teosinte domestication from Guilá Naquitz, Oaxaca. Proc. Natl. Acad. Sci. U. S. A. 98: 2104–2106.

Bockus W. W., J. A. Appel, R. L. Bowden, A. K. Fritz, B. S. Gill, *et al.*, 2001 Success Stories: Breeding for Wheat Disease Resistance in Kansas. Plant Dis. 85: 453–461.

C. elegans Sequencing Consortium, 1998 Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012–2018.

Chawla H. S., H. Lee, I. Gabur, P. Vollrath, S. Tamilselvan-Nattar-Amutha, *et al.*, 2021 Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. Plant Biotechnol. J. 19: 240–250.

Chia J.-M., C. Song, P. J. Bradbury, D. Costich, N. de Leon, *et al.*, 2012 Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44: 803–807.

Citovsky V., and I. Keren, 2017 Faculty Opinions recommendation of Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature.

Crow J. F., 1998 90 years ago: the beginning of hybrid maize. Genetics 148: 923–928.

Cui J., N. Shen, Z. Lu, G. Xu, Y. Wang, *et al.*, 2020 Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. Plant Methods 16: 85.

Della Coletta R., Y. Qiu, S. Ou, M. B. Hufford, and C. N. Hirsch, 2021 How the pan-genome is changing crop genomics and improvement. Genome Biol. 22: 3.

Dempewolf H., G. Baute, J. Anderson, B. Kilian, C. Smith, *et al.*, 2017 Past and future use of wild relatives in crop breeding. Crop Sci. 57: 1070–1082.

Dijk E. L. van, Y. Jaszczyszyn, D. Naquin, and C. Thermes, 2018 The Third Revolution in Sequencing Technology. Trends Genet. 34: 666–681.

Doebley J., A. Stec, and L. Hubbard, 1997 The evolution of apical dominance in maize. Nature 386: 485–488.

Doebley J. F., B. S. Gaut, and B. D. Smith, 2006 The Molecular Genetics of Crop Domestication. Cell 127: 1309–1321.

Dubcovsky J., and J. Dvorak, 2007 Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. Science 316: 1862–1866.

Duvick D. N., 2005 The Contribution of Breeding to Yield Advances in maize (Zea mays L.), pp. 83–145 in *Advances in Agronomy*, Academic Press.

Eid J., A. Fehr, J. Gray, K. Luong, J. Lyle, *et al.*, 2009 Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138.

Faris J. D., 2014 Wheat Domestication: Key to Agricultural Revolutions Past and Future, pp. 439–464 in *Genomics of Plant Genetic Resources: Volume 1. Managing, sequencing and mining genetic resources*, edited by Tuberosa R., Graner A., Frison E. Springer Netherlands, Dordrecht.

Flint-Garcia S. A., A.-C. Thuillet, J. Yu, G. Pressoir, S. M. Romero, *et al.*, 2005 Maize association population: a high-resolution platform for quantitative trait locus dissection. The Plant Journal 44: 1054–1064.

Gage J. L., B. Monier, A. Giri, and E. S. Buckler, 2020 Ten Years of the Maize Nested Association Mapping Population: Impact, Limitations, and Future Directions. Plant Cell 32: 2083–2093.

Garalde D. R., E. A. Snell, D. Jachimowicz, B. Sipos, J. H. Lloyd, *et al.*, 2018 Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15: 201–206.

García-Lara S., and S. O. Serna-Saldivar, 2019 Chapter 1 - Corn History and Culture, pp. 1–18 in *Corn (Third Edition)*, edited by Serna-Saldivar S. O. AACC International Press, Oxford.

Goffeau A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, *et al.*, 1996 Life with 6000 genes. Science 274: 546, 563–7.

Gonzalez-Garay M. L., 2016 Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq), pp. 141–160 in *Transcriptomics and Gene Regulation*, edited by Wu J. Springer Netherlands, Dordrecht.

Gosal S. S., and S. H. Wani, 2020 *Accelerated Plant Breeding, Volume 1: Cereal Crops*. Springer Nature.

Haberer G., N. Kamal, E. Bauer, H. Gundlach, I. Fischer, *et al.*, 2020 European maize genomes highlight intraspecies variation in repeat and gene content. Nat. Genet. 52: 950–957.

Hajjar R., and T. Hodgkin, 2007 The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica 156: 1–13.

Hake S., and J. Ross-Ibarra, 2015 Genetic, evolutionary and plant breeding insights from the domestication of maize. Elife 4. https://doi.org/10.7554/eLife.05861

He F., R. Pasam, F. Shi, S. Kant, G. Keeble-Gagnere, *et al.*, 2019 Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat. Genet. 51: 896–904.

Heather J. M., and B. Chain, 2016 The sequence of sequencers: The history of sequencing DNA. Genomics 107: 1–8.

Hirsch C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni, *et al.*, 2014 Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26: 121–135.

Hirsch C. N., C. D. Hirsch, A. B. Brohammer, M. J. Bowman, I. Soifer, *et al.*, 2016 Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. Plant Cell 28: 2700–2714.

Hu Y., V. Colantonio, B. S. F. Müller, K. A. Leach, A. Nanni, *et al.*, 2021 Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. Nat. Commun. 12: 1227.

Hufford M. B., X. Xu, J. van Heerwaarden, T. Pyhäjärvi, J.-M. Chia, *et al.*, 2012 Comparative population genomics of maize domestication and improvement. Nat. Genet. 44: 808–811.

Hufford M. B., A. S. Seetharam, and M. R. Woodhouse, 2021 De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. bioRxiv.

International Human Genome Sequencing Consortium, 2004 Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.

International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators:, R. Appels, K. Eversole, C. Feuillet, *et al.*, 2018 Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361. https://doi.org/10.1126/science.aar7191

Jain M., H. E. Olsen, B. Paten, and M. Akeson, 2016 Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17: 256.

Jayakodi M., M. Schreiber, N. Stein, and M. Mascher, 2021 Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Res. 28. https://doi.org/10.1093/dnares/dsaa030

Jiao Y., H. Zhao, L. Ren, W. Song, B. Zeng, *et al.*, 2012 Genome-wide genetic changes during modern breeding of maize. Nat. Genet. 44: 812–815.

Jiao Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, *et al.*, 2017 Improved maize reference genome with single-molecule technologies. Nature 546: 524–527.

Jung H., C. Winefield, A. Bombarely, P. Prentis, and P. Waterhouse, 2019 Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. Trends Plant Sci. 24: 700–724.

Khoury C. K., D. Carver, S. L. Greene, K. A. Williams, H. A. Achicanoy, *et al.*, 2020 Crop wild relatives of the United States require urgent conservation action. Proc. Natl. Acad. Sci. U. S. A. 117: 33351–33357.

Kolmer J. A., A. Bernardo, G. Bai, M. J. Hayden, and S. Chao, 2018a Adult Plant Leaf Rust Resistance Derived from Toropi Wheat is Conditioned by Lr78 and Three Minor QTL. Phytopathology 108: 246–253.

Kolmer J. A., Z. Su, A. Bernardo, G. Bai, and S. Chao, 2018b Mapping and characterization of the new adult plant leaf rust resistance gene Lr77 derived from Santa Fe winter wheat. Theor. Appl. Genet. 131: 1553–1560.

Kumar S., S. C. Bhardwaj, O. P. Gangwar, A. Sharma, N. Qureshi, *et al.*, 2021 Lr80: A new and widely effective source of leaf rust resistance of wheat for enhancing diversity of resistance among modern cultivars. Theor. Appl. Genet. 134: 849–858.

Li L., S. R. Eichten, R. Shimizu, K. Petsch, C.-T. Yeh, *et al.*, 2014 Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol. 15: R40.

Li Q., J. I. Gent, G. Zynda, J. Song, I. Makarevitch, *et al.*, 2015 RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc. Natl. Acad. Sci. U. S. A. 112: 14728–14733.

Liu J., A. R. Fernie, and J. Yan, 2020 The Past, Present, and Future of Maize Improvement: Domestication, Genomics, and Functional Genomic Routes toward Crop Enhancement. Plant Commun 1: 100010.

Luo M.-C., Y. Q. Gu, D. Puiu, H. Wang, S. O. Twardziok, *et al.*, 2017 Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature 551: 498–502.

Maccaferri M., N. S. Harris, S. O. Twardziok, R. K. Pasam, H. Gundlach, *et al.*, 2019 Durum wheat genome highlights past domestication signatures and future improvement targets. Nat. Genet. 51: 885–895.

Mao H., H. Wang, S. Liu, Z. Li, X. Yang, *et al.*, 2015 A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nat. Commun. 6: 8326.

Matsuoka Y., Y. Vigouroux, M. M. Goodman, J. Sanchez G, E. Buckler, *et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. U. S. A. 99: 6080–6084.

McCarty D. R., A. M. Settles, M. Suzuki, B. C. Tan, S. Latshaw, *et al.*, 2005 Steady-state transposon mutagenesis in inbred maize: Maize steady-state transposon mutagenesis. Plant J. 44: 52–61.

McFadden E. S., and E. R. Sears, 1946 The origin of Triticum spelta and its free-threshing hexaploid relatives. J. Hered. 37: 81–89.

McMullen M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. Li, *et al.*, 2009 Genetic properties of the maize nested association mapping population. Science 325: 737–740.

Mikel M. A., and J. W. Dudley, 2006 Evolution of north American dent corn from public to proprietary germplasm. Crop Sci. 46: 1193–1205.

Mikel M. A., 2011 Genetic composition of contemporary U.s. commercial dent corn germplasm. Crop Sci. 51: 592–599.

Minio A., M. Massonnet, R. Figueroa-Balderas, A. M. Vondras, B. Blanco-Ulate, *et al.*, Iso-Seq allows genome-independent transcriptome profiling of grape berry development. G3: Genes|Genomes|Genetics 9: 755 LP–767.

Nannas N. J., and R. K. Dawe, 2015 Genetic and genomic toolbox of Zea mays. Genetics 199: 655–669.

Park R. F., 2016 Wheat: Biotrophic Pathogen Resistance. Encyclopedia of Food Grains 264–272.

Parker M. T., K. Knop, A. V. Sherwood, N. J. Schurch, K. Mackinnon, *et al.*, 2020 Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. Elife 9. https://doi.org/10.7554/eLife.49658

Pont C., T. Leroy, M. Seidel, A. Tondelli, W. Duchemin, *et al.*, 2019 Tracing the ancestry of modern bread wheats. Nat. Genet. 51: 905–911.

Prasanna B. M., 2012 Diversity in global maize germplasm: characterization and utilization. J. Biosci. 37: 843–855.

Qaim M., 2020 Role of new plant breeding technologies for food security and sustainable agricultural development. Appl. Econ. Perspect. Policy 42: 129–150.

Qureshi N., H. Bariana, V. V. Kumran, S. Muruga, K. L. Forrest, *et al.*, 2018 A new leaf rust resistance gene Lr79 mapped in chromosome 3BL from the durum wheat landrace Aus26582. Theor. Appl. Genet. 131: 1091–1098.

Romay M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts, *et al.*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14: R55.

Schnable P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, *et al.*, 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

Shendure J., S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, *et al.*, 2017 DNA sequencing at 40: past, present and future. Nature 550: 345–353.

Shiferaw B., B. M. Prasanna, J. Hellin, and M. Bänziger, 2011 Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. Food Security 3: 307.

Shiferaw B., M. Smale, H.-J. Braun, E. Duveiller, M. Reynolds, *et al.*, 2013 Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. Food Security 5: 291–317.

Singla J., L. Lüthi, T. Wicker, U. Bansal, S. G. Krattinger, *et al.*, 2017 Characterization of Lr75: a partial, broad-spectrum leaf rust resistance gene in wheat. Theor. Appl. Genet. 130: 1–12.

Song J.-M., Z. Guan, J. Hu, C. Guo, Z. Yang, *et al.*, 2020 Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants 6: 34–45.

Springer N. M., K. Ying, Y. Fu, T. Ji, C.-T. Yeh, *et al.*, 2009 Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5: e1000734.

Springer N. M., S. N. Anderson, C. M. Andorf, K. R. Ahern, F. Bai, *et al.*, 2018 The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat. Genet. 50: 1282–1288.

Stark R., M. Grzelak, and J. Hadfield, 2019 RNA sequencing: the teenage years. Nat. Rev. Genet. 20: 631–656.

Strable J., and M. J. Scanlon, 2009 Maize (Zea mays): a model organism for basic and applied research in plant biology. Cold Spring Harb. Protoc. 2009: db.emo132.

Sun S., Y. Zhou, J. Chen, J. Shi, H. Zhao, *et al.*, 2018 Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat. Genet. 50: 1289–1295.

Sun X., S. Zhu, N. Li, Y. Cheng, J. Zhao, *et al.*, 2020 A Chromosome-Level Genome Assembly of Garlic (Allium sativum) Provides Insights into Genome Evolution and Allicin Biosynthesis. Mol. Plant 13: 1328–1339.

Technologies to boost breeding, 2018 Nat Plants 4: 1.

Tester M., and P. Langridge, 2010 Breeding technologies to increase crop production in a changing world. Science 327: 818–822.

Voss-Fels K. P., A. Stahl, B. Wittkop, C. Lichthardt, S. Nagler, *et al.*, 2019 Breeding improves wheat productivity under contrasting agrochemical input levels. Nat Plants 5: 706–714.

Walkowiak S., L. Gao, C. Monat, G. Haberer, M. T. Kassa, *et al.*, 2020 Multiple wheat genomes reveal global variation in modern breeding. Nature 588: 277–283.

Wang H., T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigouroux, *et al.*, 2005 The origin of the naked grains of maize. Nature 436: 714–719.

Watson A., S. Ghosh, M. J. Williams, W. S. Cuddy, J. Simmonds, *et al.*, 2018 Speed breeding is a powerful tool to accelerate crop research and breeding. Nat Plants 4: 23–29.

Wenger A. M., P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, *et al.*, 2019 Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37: 1155–1162.

Wulff B. B. H., and M. J. Moscou, 2014 Strategies for transferring resistance into wheat: from wide crosses to GM cassettes. Front. Plant Sci. 5: 692.

Yang N., J. Liu, Q. Gao, S. Gui, L. Chen, *et al.*, 2019 Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat. Genet. 51: 1052–1059.

Chapter 2 - Chromosome-level Genome Assembly of a Regenerable Maize Inbred Line A188 *

Abstract

The inbred line A188 is an attractive model genotype for elucidation of maize gene function and transformability due to the high embryogenic and transformable character of the line. A188 exhibits many contrasting traits to the inbred line B73, the line used for first maize reference genome. The lack of a genome sequence of A188 limits the use of A188 as a model genotype for functional studies. Here, a chromosome-level genome assembly of A188 was constructed using long reads and optical maps. Genome comparison between A188 and B73 based on both whole genome alignments and read depths from sequencing reads identified approximately 1.1 Gb syntenic sequences as well as extensive structural variation, including a 1.8 Mb duplication containing the *Gametophyte factor1* locus for unilateral cross-incompatibility and six inversions of 0.7 Mb or greater. Increased copy number of a gene, *carotenoid cleavage dioxygenase 1* (*ccd1*) in A188 compared to B73 is associated with elevated expression during seed development. High *ccd1* expression in seeds together with low expression of *yellow endosperm 1* (*y1*) condition reduced carotenoid accumulation, which accounts for the white seed phenotype of A188 that contrasts with the yellow seed of B73 that has high expression of *y1* and low expression of the single-copy *ccd1*. Further, transcriptome and epigenome analyses with the A188 reference genome revealed enhanced expression of defense pathways and altered DNA methylation patterns of embryonic callus. The A188 genome assembly provides a high-resolution sequence for a complex

* Reprinted from Lin G., C. He, J. Zheng, D.-H. Koo, H. Le, *et al.*, 2020 Chromosome-level Genome Assembly of a Regenerable Maize Inbred Line A188. bioRxiv 2020.09.09.289611.

genome species and a foundational resource for analyses of genome variation and gene function in maize. The genome, in comparison to B73, contains extensive intra-species structural variations and other genetic differences. Expression and network analyses identified discrete profiles for embryonic callus and other tissues.

Introduction

The maize inbred line A188 was derived from a line related to the commercial maize variety Silver King and a northwestern dent line (Gerdes *et al.* 1993). A188 is amenable to somatic embryogenic culture (**Figure B.1**) and regeneration and was the first maize line used to produce genetically modified plants (Rhodes *et al.* 1988). A popular maize transformation line, Hi-II, was isolated from offspring of a cross between A188 and B73, an elite maize reference inbred line (Armstrong *et al.* 1991; Vega *et al.* 2008). Although highly valuable for plant regeneration and transformation, A188 is not agronomic competitive, having small ears and low grain yield (**Figure 2.1, Table B.1**). The line also exhibits a high degree of growth habit plasticity in response to varying environments. In particular, A188 is overly sensitive to abiotic and biotic stresses, including drought, heat, and bacterial and fungal diseases, in comparison to elite maize lines (Wisser *et al.* 2011). Nonetheless, hybrids of A188 and B73 exhibit extensive heterosis (**Figure B.2**). A188, therefore, in addition to traits related to transformability, can serve as a model inbred line for the genetic dissection of many important agronomic traits, heterosis, and plant-environment interactions.

Efforts have been pursued to develop efficiency and quality strategies for maize genome sequencing and assemblies. The first maize reference genome for B73 was sequenced and assembled using bacterial artificial chromosomes (BACs) (Schnable *et al.* 2009). Since then,

additional assemblies have been produced using so-called next generation high throughput sequencing, including both short and long read technologies (Hirsch *et al.* 2016; Jiao *et al.* 2017; Sun *et al.* 2018; Springer *et al.* 2018; Yang *et al.* 2019; Ou *et al.* 2020; Haberer *et al.* 2020). Recently, two long-read technologies, PacBio and Nanopore, were combined with optical DNA mapping to produce a high-continuity maize assembly (Liu *et al.* 2020). Here, Nanopore long reads and optical DNA mapping were used to construct a chromosome-level maize genome of A188 for discovery of structural variation as well as performed transcriptomes and DNA methylome analyses of embryogenic callus.

Materials and Methods

Genetic materials

A188 (PI 693339) seeds were obtained from the North Central Regional Plant Introduction Station in Ames, IA. The A188 inbred line was derived from a cross between the inbred line 4-29 and the inbred line 64, also named A48, followed by four generations of backcross with 4-29. The line 4-29 was derived from the commercial variety Silver King and the line 64 was from a northwestern dent line (Gerdes *et al.* 1993). Double haploid lines were developed from the F_1 of B73 (PI 550473) x A188 at the Doubled Haploid Facility at Iowa State University.

Nanopore A188 whole-genome sequencing

A188 were grown in the greenhouse at 28°C and 23°C day/night, with a photoperiod of 14:10 h (light:dark). Nuclei were isolated from seedling leaves using a modified nucleus isolation protocol (Zhang *et al.* 2012) and dissolved in buffer G2 (Qiagen). The lysate was used for DNA isolation with Qiagen DNeasy Plant Mini Kit (Qiagen) following the manufacturer protocol. A188 genomic DNA was size selected for 15-30 kb and above with the BluePippin cassette kit BLF7510 (Sage Science) with high-pass-filtering protocol, followed by a library preparation with the SQK-

LSK109 kit (Oxford Nanopore). Each DNA library was loaded on a FLO-MIN106D flowcell and sequenced on MinION (Oxford Nanopore). FAST5 raw data were converted to FASTQ data using the basecaller Guppy (version 3.4.4, Oxford Nanopore) with default parameters.

Illumina A188 whole-genome sequencing

Three independent A188 leaf samples were collected for extracting nuclear DNAs. Two were used for PCR-free paired-end 2x125 bp Illumina sequencing and one was used for PCR-free paired-end 2x250 bp Illumina sequencing on Hiseq2500 at Novogene. In addition, genomic DNA was extracted from A188 immature ears for additional PCR-free paired-end 2x250 bp Illumina sequencing. Therefore, comparable 2x250 bp data were generated from the leaf and ear tissue samples. The 2x125 bp Illumina sequencing data were comparable with the previously generated 2x125 bp B73 whole genome sequencing data (SRR4039069 and SRR4039070) (Liu *et al.* 2017a), both of which were used for Comparative Genomic Read Depth (CGRD) analysis.

Assembly of Nanopore data *via* Canu

FASTQ Nanopore data were assembled with Canu (1.9) with the following options: "'corMhapOptions=--threshold 0.8 --ordered-sketch-size 1000 --ordered-kmer-size 14' correctedErrorRate=0.105 genomeSize=2.4g minReadLength=10000 minOverlapLength=800 corOutCoverage=60".

Contigs filtering

Leaf and ear 2x250 bp data were aligned to the contigs with the “mem” module in bwa (0.7.12-r1039) (Li and Durbin 2010). Uniquely mapped reads with less than 15% mismatches were used to determine read counts per contig with the “intersect” module of BEDTools (v2.29.2) (Quinlan and Hall 2010). The log2 of the ratio of read counts normalized by using total reads of leaf and ear samples was calculated for each contig. The contigs with a log2 value larger than 0.5

were considered as the contigs with variable counts from leaf and ear samples. The contigs (N=21) that had variable counts and less than 100 kb and were not anchored to B73Ref4 via Ragoo (version 1.2) (Alonge *et al.* 2019) were discarded. In addition, the contigs (N=16) smaller than 15 kb were also discarded.

Through analysis of read counts, the contigs that had variable counts and matched with the previously sequenced mitochondrial genome sequence (Genbank accession: DQ490952.1) and the chloroplast genome sequence (Genbank accession: KF241980.1) were identified. One chloroplast contig and 13 mitochondrion contigs were found. The chloroplast contig had almost identical sequences to the Genbank accession KF241980.1. The failure of assembling mitochondrial contigs into one was likely due to heterogeneous forms of mitochondria. In the A188 genome assembly, A188Ref1, the previously assembled, DQ490952.1 and KF241980.1, were used to represent the mitochondrion and chloroplast genomes, respectively.

Sequence polishing of assembled contigs

After filtering contigs from organelles or contamination, the remaining contigs were first polished with raw Nanopore reads that contained signal information using Nanopolish (0.11.0) (github.com/jts/nanopolish). Briefly, Nanopore reads were aligned with the contigs using the aligner Minimap2 (2.14-r892) (Li 2016). Polymorphisms, including small insertions and deletions as well as single nucleotide polymorphisms, were called and sequence errors were corrected. The Nanopolish polishing was performed twice, followed by twice polishing with Illumina sequencing data using Pilon (version 1.23) (Walker *et al.* 2014). In each Pilon polishing, reads were aligned to contigs with the module of "mem" in bwa (0.7.12-r1039) (Li and Durbin 2010). Contigs were corrected with the parameters of "--minmq 40 --minqual 15" using Pilon.

Hybrid scaffolding with Bionano data and polished contigs

Bionano raw molecules were filtered to remove molecules less than 100 kb. The remaining molecules were assembled into Bionano maps with the assembly module in the software Bionano Tools (v1.0). Five times extension and merge iterations and noise parameters were automatically determined by using the parameters of "-i 5 -y". The hybrid scaffolding module from the Bionano Tools was used for scaffolding polished contigs. The conflict filter level for both genome maps and sequences were set to 2 by using the parameters of "-B 2 -N 2".

Construction of a B73xA188 genetic map

Genomic DNA of DH lines was extracted by using BioSprint 96 DNA Plant Kit (Qiagen), and normalized to 10 ng/uL for tunable Genotyping-By-Sequencing (GBS) modified from tunable GBS (Ott *et al.* 2017). Briefly, for each genomic DNA sample, the restriction enzyme *Bsp1286I* (NEB) was used for DNA digestion for 3 h at 37^{o}C, followed by ligation with a barcoded single-stranded oligo with T4 DNA ligase (NEB) for 1 h at 16^{o}C. Enzymatic activity was inactivated at 65^{o}C for 20 min and all samples of ligated DNA were pooled, followed by purification with Qiagen PCR purification kit (Qiagen). The purified ligated DNA was subjected to PCR amplification with Q5 high-fidelity DNA polymerase (NEB), followed by purification with Agencourt AMPure XP (Beckman Coulter). The final sequencing library product was prepared by size selection at the range of 200 to 400 bp by a Pippin Prep run with 2% agarose gel cassettes (Sage Science). Illumina sequencing was performed on a HiseqX 10 at Novogene (USA).

Raw FASTQ data were deconvoluted to multiple samples and trimmed to remove barcode sequences and low-quality bases using Trimmomatic (version 0.38). Clean reads were aligned to polished contigs with the "mem" module of bwa and uniquely mapped reads with less than 8% mismatches were used for SNP analysis. SNPs were discovered by HaplotypeCaller of GATK (version 4.1.0.0) and filtered by SelectVariants of GATK to select biallelic variants (McKenna *et*

al. 2010). SNP sites with at most 80% missing data, at least 10% minor allele frequency and at most 5% heterozygous rates remained. A segmentation (or binning) algorithm was implemented to determine genotypes of chromosomal segments in each DH line (Ott *et al.* 2017). Genotypes of bin markers of 100 DH lines were used to construct a genetic map with MSTmap (Wu *et al.* 2008).

Another genetic map (**Supplemental_file_2.11**) was built using A188Ref1 as the reference genome with 37 additional DH lines. Recombination data was inferred from the genetic map (BAgm.v02) based on A188Ref1.

ALLMaps to build pseudomolecules

The genetic map that was built based on polished contigs as the reference genome was used for further scaffolding. Each scaffold harbored more than 10 markers. In total, 29 scaffolds were on the map. Scaffolds were aligned to B73Ref4 via NUCMer. Based on the orientation of scaffolds relative to B73Ref4 chromosomes, the order of markers in each linkage group was either kept the same order or flipped to match their orders in B73Ref4. The software ALLMaps (JCVI utility libraries v1.0.6) (Tang *et al.* 2015) was conducted with default parameters, constructing 10 pseudomolecules corresponding to ten A188 chromosomes.

BUSCO assessment

Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão *et al.* 2015) was run in a mode of "genome" to assess the completeness of the assembly with default parameters. BUSCO was run in a mode of "transcriptome" to assess the completeness of the gene annotation with default parameters. Both assessments using the Liliopsida database (liliopsida_odb10) that consisted of 3,278 conserve core genes.

Estimation of base errors using KAD analysis

The module "KADprofile.pl" in the KAD tool (version 0.1.7) (He *et al.* 2020) was used to estimate errors in A188Ref1. The input read data were the merged trimmed Illumina 2x250 bp reads from leaf and immature ears. The k-mer length of 47 mer was used.

Estimation of recombination rates

Genetic distances of non-overlapping 1-Mb windows was estimated. Non-overlapping 1 Mb windows were generated by the module of "makewindows" in BEDTools (v2.29.2) (Quinlan and Hall 2010). The last window of each chromosome was discarded due to the smaller size than 1 Mb. The prediction of genetic distance per window utilized a method developed previously (Liu *et al.* 2009). Briefly, a generalized additive model (GAM) was used for the prediction of the genetic distance of any physical interval.

The similar method was used to estimate recombination rates around each gene and repetitive element. For example, for a given element, we first find the midpoint of the element. The genetic positions were then predicted, by GAM, for the position 0.5 Mb upstream and the position 0.5 Mb downstream. The distance of the genetic positions was then used to represent the recombination context of the element.

The recombination rates that are lower than 0.6 cM/Mb and higher than 3 cM/Mb were categorized to low recombination and high recombination, respectively.

Callus induction from immature embryos

A188 ears were harvested at 11 days after pollination (DAP11), and surface-sterilized for 30 minutes in 50% (v/v) bleach (6% sodium hypochlorite) that contains 3-4 drops of Tween 20 followed by three washes in sterile distilled water. Immature embryos of size 1.0-1.5 mm were isolated and cultured on callus induction medium (CIM) (Rakshit *et al.* 2010). CIM was composed of Chu N6 basal medium with vitamins (Chu *et al.* 1975) supplemented with 2.3 g/L L-proline,

200 mg/L casein hydrolysate, 3% sucrose, 1 mg/L 2,4-dichlorophenoxyacetic acid, 3 g/L gelrite, pH 5.8. Subculture was conducted every 14 days. The 39-days callus samples were collected for methylome and transcriptome analysis.

Illumina RNA-Seq, transcriptome assembly, and differential expression

Thirty-three RNA samples were extracted from 11 diverse tissue types of A188 with three biological replicates using RNeasy Plant Mini Kit (Qiagen) (**Table B.2**). Briefly, the 11 tissues included the root and the above-ground of 10-day-old seedling, three different parts of the 11th leaf tissue at V12, the meiotic tassel, anther, and immature ear at V18, the endosperm and embryo 16 days after pollination, and the callus after 39 days culture of DAP11 immature embryos. RNA quality control, library preparation, and sequencing were performed on an Illumina Novaseq 6000 platform at Novogene (USA). Trimmomatic (version 0.38) (Bolger *et al.* 2014) was used to trim the adaptor sequence and low-quality bases of RNA-Seq raw reads. The parameters used for the trimming is "ILLUMINACLIP:trimming_db:3:20:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:13 MINLEN:40". The trimming adaptor database (trimming_db) includes the sequences: adaptor1, TACACTCTTTCCCTACACGACGCTCTTCCGATCT; adaptor2, GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT. Only these trimmed paired reads both of which were at least 40 bp after trimming were retained for further analysis.

Trimmed reads were aligned to A188 (A188Ref1) using HISAT2 (version 2.1.0) with the parameters of "-p 8 --dta --no-mixed --no-discordant -k 5 -x" (Kim *et al.* 2019). Alignments whose paired reads were concordantly paired were kept. The software StringTie2 (version 2.1.0) (Kovaka *et al.* 2019) was used to assemble the transcriptome with alignments from a dataset of each A188 sample with the default parameters. In total, 33 transcriptome assemblies from 33 samples were

generated. All transcriptome assemblies were merged to build an A188 Illumina transcriptome assembly using the merge function in StringTie2.

Differential expression of the callus relative to other tissue types

Trimmed reads were aligned to A188Ref1 with STAR (2.7.3a) (Dobin *et al.* 2013). Uniquely mapped reads with at least 96% coverage and 96% identity were used for determining read counts per gene. DESeq2 (version 1.26.0) was used to identify differential expression between the callus and each of other tissue types. Multiple tests were corrected with the FDR (false discovery rate) approach (Benjamini and Hochberg 1995). The FDR of 5% was set as the threshold.

Nanopore A188 cDNA direct sequencing

The same three biological replicates of the seedling and callus samples used for Illumina RNA-Seq were sequenced using the Nanopore direct cDNA sequencing protocol. Briefly, mRNA was first isolated from 10 ug total RNA with Poly(A) RNA Selection Kit (Lexogen), followed by direct cDNA library preparation with SQK-DCS109 kit (Oxford Nanopore) using the protocol version DCS_9090_v109_revB_04Feb2019. The cDNA library was loaded onto a FLO-MIN106D R9 flowcell and sequenced on MinION (Oxford Nanopore).

FAST5 raw data was converted to FASTQ data using the basecaller Guppy version 3.4.5 (Oxford Nanopore) with default parameters. Two trimming steps were employed. Adapter sequence was first trimmed by porechop (version 0.2.4) (https://github.com/rrwick/Porechop) with parameters "--check_reads 10000 --adapter_threshold 100 --end_size 100 --min_trim_size 5 --end_threshold 80 --extra_end_trim 1 --middle_threshold 100 --extra_middle_trim_good_side 5 --extra_middle_trim_bad_side 50", and then poly A was trimmed by the software cutadapt (version 2.6) (Martin 2011) with the options of " -g T{12} -e 0.1 -a A{12} -n 100". Trimmed reads were

aligned to A188Ref1 as unstranded spliced long reads using MiniMap2 (version 2.14) (Li 2018) with the parameter "-ax splice". Merged alignments from three replicates were input to StingTie2 for generating assembled transcripts.

Genome annotation

The Maker (2.31.10) was used for genome annotation (Holt and Yandell 2011). The genome was masked by using Repeatmasker (4.0.7) (Smit *et al.* 2013-2015) with the A188 repeat library built by the Extensive *de novo* TE Annotator (EDTA, v1.8.4) (Ou *et al.* 2019). Two rounds of the maker prediction were performed. At the first round, the A188 assembled transcripts and B73Ref4 protein data were used as EST and protein evidence, respectively. The parameters "est2genome=1" and "protein2genome=1" were set to directly produce gene models from transcripts and proteins. At this round, no *ab initio* gene predictors were used. Prior to the second maker round, a snap model was trained using the confident gene set from the first round. Gene models produced from round 1 were input as one of predicted gene models. These gene models were competed with gene models predicted by three gene predictors: snap (2013_11_29) (Korf 2004), augustus (3.3.3) (Stanke and Waack 2003), and fgenesh (v.8.0.0) (softberry.com). ESTs from relative maize genotypes and proteins from closely related species were provided as additional evidence. Gene models output from Maker were further filtered. First genes matched with the following criteria "-evalue 1e-50 -qcov_hsp_perc 60" to the transposon database in Maker were filtered. Second, a transcript retained if it carried Pfam domains from the result of InterProScan (version 5.39-77.0) and/or had an annotation edit distance (AED) less than 0.4, which measured the level of discrepancy of an annotation from supporting evidence.

Functional annotation of transcripts

BLASTP was used to map all proteins to the SWISS-Prot database (https://www.uniprot.org/) with the e-value cutoff of 1e-6. Gene ontology (GO) was extracted from InterProScan.

Identification of a major transcript per gene

For a gene containing multiple transcripts, a major transcript per gene was selected if a transcript had the highest non-zero FPKM (Fragments Per Kilobase of transcript per Million mapped reads) determined from Illumina RNA-Seq datasets of diverse tissues by Cufflink (v2.2.1) (Trapnell *et al.* 2010), and/or the lowest BLASTP e-value to the SWISS-Prot database, and/or the longest transcript length. The BLASTP e-value had a priority relative to the transcript length. If data were not sufficient to make a decision, the one with the longest length was selected.

Syntenic genes between A188 and B73

Syntenic genes were identified with MCscan (JCVI utility libraries v1.0.6) (Tang *et al.* 2008). Major transcripts were used as the input and the parameter "--cscore=.99" was used to find 1-to-1 syntenic gene relationship.

Paralogs in A188 and orthologs between A188 and B73

Paralogs in A188 and orthologs between A188 and B73 were identified with OrthoMCL (Li *et al.* 2003). Briefly, protein sequences of major transcripts with at least 20 amino acids were used for all-to-all BLASTP with the e-value cutoff of 1e-5. The BLASTP result was input to OrthoMCL to identify paralogous and orthologous groups.

Identification of gene clusters

A gene cluster was defined if at least three genes from a group of A188 paralogs identified by OrthoMCL were physically closely located on a chromosome. The maximum distance is 250 kb for two neighboring genes in a cluster.

Annotation of NLR genes

The NLR genes of A188Ref1 were annotated using the NLR-Annotator pipeline (Steuernagel *et al.* 2020).

Repeat annotation

EDTA (v1.8.4) (Ou *et al.* 2019) was used for repeat annotation with, maize as the "species" input, the curated maize transposable element database from https://github.com/oushujun/MTEC as the "curatedlib" input, and B73 coding sequences as the "cds' input.

Analysis of NUMT and NUPT

The "nucmer" command from the software MUMmer 4 (Kurtz *et al.* 2004) was used to align the A188 mitochondria or chloroplast genomes to A188Ref1. For mitochondrial alignments, each required at least 5 kb and 95% identity. For chloroplast DNA alignments, each required at least 3 kb and 95% identity based on the minimal requirement for a sufficient Fluorescence *in situ* hybridization **(**FISH) signal (Roark *et al.* 2010). Multiple alignments with the distance less than 100 kb were clustered into a block, considered to be a nuclear integration event.

Comparative genomic analysis via SyRI and CGRD

The "nucmer" command was used for whole genome alignment of 10 chromosomal pseudomolecules between A188Ref1 and B73Ref4. The parameter of "--maxmatch -c 500 -b 500 -l 50" was used in the command "nucmer" and the parameter of "-i 95 -l 1000 -m" in the command of delta-filter, which resulted in best alignments with at least 1 kb matches and at least 95% identity between the two assembled genomes. The "show-coords" command with the parameter of "-THrd" was run to convert alignments to a tab-delimited flat text format. Alignment results were then used for identifying genomic structural variation and nucleotide polymorphisms through SyRI (v1.2) (Goel *et al.* 2019) with the parameter of "--allow-offset 100". SyRI analysis discovered genome

duplication, translocation, inversion, as well as syntenic, unaligned, divergent sequences. SNPs, small insertions, and deletions were identified as well.

The CGRD pipeline (v0.1) (github.com/liu3zhenlab/CGRD) was employed to find copy number variation (CNV) through comparing depths of Illumina reads from A188 and B73 with the default parameters (Peng *et al.* 2019). A value of the log2 read depth ratio per sequence segment ($LogRD$) is the indication for CNV. For a segment, the $LogRD$ is close to zero if sequences of two genotypes are identical and no CNVs. The sufficient derivation of the mean of $LogRD$ from zero is likely due to CNV or a high level of divergence. CGRD was performed using A188Ref1 as the reference genome and identified sequences of A188Ref1 showing conserved (B73=A188), copy number plus (B73>A188), and copy number minus (B73<A188) in B73 relative to A188. When B73Ref4 was used as the reference, the analysis found sequences of B73Ref4 showing conserved (A188=B73), copy number plus (A188>B73), and copy number minus (A188<B73) in A188 relative to B73.

Identification of presence and absence variance (PAV) or highly divergent sequences (HDS)

SyRI analysis listed B73Ref4 sequences that were not aligned to A188Ref1, and *vice versa*, as well as insertion/deletion polymorphisms between the two chromosomal sequences. Unaligned sequences or insertion/deletion polymorphisms identified by SyRI were compared with CGRD segments. For each SyRI event, a supporting score of read depth data from CGRD was determined by using the formula of $\sum_i^n \frac{-LogRD_i \times O_i}{L}$, where i represents the ith overlap between a CGRD segment and a SyRI event; $LogRD$ stands for the $LogRD$ of the CGRD segment and only negative values were taken into calculation; O is the overlapping length in bp; L is the length in bp of the SyRI event; and n is the total number of overlaps. The resulting value from the formula represents

the degree of the differentiation in read depth between the two genotypes for the SyRI event. The higher the number, the more confidence the PAV or HDS event. A SyRI event is considered to be a PAV or HDS if a supporting score is larger than 3.

Identification of large inversion events

Inversion, from kb to Mb levels of events, between A188Ref1 and B73Ref4 were revealed by SyRI. Large events with both A188 and B73 sequences larger than 0.5 Mb were exacted. First, the inversion sequences of B73Ref4 were aligned to B73Ref5 to confirm the inverted orientation relative to A188Ref1. For a given inversion, if >80% B73Ref4 sequences were aligned to B73Ref5 in the plus orientation, the inversion was supported by B73Ref5. If <20% B73Ref4 sequences aligned to B73Ref5 were in the plus orientation, the inversion was considered to not be supported by B73Ref5. Second, the recombination frequency between the start and the end of an inversion event was estimated and adjusted to cM per Mb. Third, SNPs between the two genomes and located on the inversion were identified. The common SNPs genotyped in the maize 282 (Bukowski *et al.* 2018) population were extracted for determining linkage disequilibrium (LD) between SNPs in distance of 0.2-0.3 Mb. Vcftools (v0.1.17) (Danecek *et al.* 2011) was employed to calculate LD. The genome-wide LDs between SNPs in distance of 0.2 Mb were determined as the control.

Structure analysis of inversions in maize HapMap2 population

The software STRUCTURE (v2.3.4) (Pritchard *et al.* 2000) was used to analyze the inference of population structure for A188 inversions in maize HapMap2 population (Hufford *et al.* 2012). A188 and B73 SNPs between inversion regions were discovered by SyRI. HapMap2 genotyping data overlapping with inversion SNPs were extracted and the subset SNPs with the missing rate less than 20% were input for STRUCTURE analysis. The major alleles, minor allele and missing locus in SNP dataset were converted to 0, 1, and -1, respectively. K=2 as the cluster

number and 10 replicate runs of the admixture model were used, with a burn-in of 10,000 iterations and a run length of 20,000 steps.

Fluorescence *in situ* hybridization (FISH)

Mitotic and meiotic chromosomes were prepared as described by Koo and Jiang (2009) with minor modifications (Koo and Jiang 2009). Root tips were collected from seedling plants and treated in a nitrous oxide gas chamber for 1.5 h, fixed overnight in ethanol:glacial acetic acid (3:1), and then squashed in a drop of 45% acetic acid. Anthers were squashed in 45% acetic acid on a slide and checked under a phase microscope. All preparations were stored at -70°C until use.

DNA probes of the CentC, Knob, Cent4 (Koo *et al.* 2016), and the probes for examining NUMTs, the PME cluster, and a potential large inversion on chromosome 4 (**Table B.3**) were labeled with digoxigenin-11-dUTP (Roche, Indianapolis, IN), biotin-16-dUTP (Roche), and/or DNP-11-dUTP (PerkinElmer), depending on whether two or three probes were used in the FISH experiment (Koo *et al.* 2016). The FISH hybridization procedure was according to a previously published protocol (Koo *et al.* 2011). After post-hybridization washes, the probes were detected with Alexafluor 488 streptavidin (Invitrogen) for biotin-labeled probes, and rhodamine-conjugated anti-digoxigenin for dig-labeled probe (Roche). The DNP-labeled probe was detected with rabbit anti-DNP, followed by amplification with a chicken anti-rabbit Alexafluor 647 antibody (Invitrogen). Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield antifade solution (Vector Laboratories). The images were captured with a Zeiss Axioplan 2 microscope (Carl Zeiss Microscopy LLC) using a cooled CCD camera CoolSNAP HQ2 (Photometrics) and AxioVision 4.8 software. The final contrast of the images was processed using Adobe Photoshop CS5 software (Adobe).

QTL mapping

Kernel colors of 125 B73xA188 DH lines were scored as 1 to 6 (1=white, 6=yellow, and 2 to 5 indicated colors between white and yellow). QTL mapping of kernel color was performed by using scanone function with Haley-Knott regression method in the R package rqtl (Broman *et al.* 2003). The LOD cutoff was the 5% highest LOD value from 1,000 permutations of phenotypic data.

qRT-PCR

qRT-PCR was used to measure the gene expression of *ccd1* and *y1* gene in genotypes of A188 and B73 as well as two DH lines DH305 and DH312. Immature ears of the four genotypes were harvested from the summer nursery in Manhattan Kansas at 16 days after pollination (DAP16). Fifteen kernels were randomly sampled from the middle of an ear, five kernels of which were pooled as a biological replication for RNA isolation. cDNA was synthesized with Verso cDNA Kit (Thermo Scientific) following the manufacturer's protocol. qRT-PCR was performed in a reaction of 10 ul with the IQTM SYBR Green Supermix reagent (BioRad) on the CFX96 Real-Time PCR System (BioRad). The thermocycling conditions for the PCR included an initial denaturation at 95˚C for 3 minutes, followed by 40 cycles of denature at 95˚C for 15 seconds, annealing and extension at 60˚C for 30 seconds. The housekeeping reference gene *actin1* was used as the internal control. Cycle threshold values (*Ct*) of two technical replicates were averaged and used to quantify relative gene expression levels. The relative expression levels of each of *ccd1* and *y1* genes in each sample were calculated using the formula $100 \times 2^{actinCt-geneCt}$, where *actinCt* and *geneCt* stand for the *Ct* values of *actin1* and *ccd1* (or *y1*), respectively. The primers used are as follows: *actin1*: act1_qrt_2F and act1_qrt_2R; *ccd1*: ccd1_qrt_5F and ccd1_qrt_5R; *y1*: y1_qrt_4F and y1_qrt_4R(**Table B.3**).

Whole-genome bisulfite sequencing (WGBS)

DNA from seedling and the callus samples that were used for RNA-Seq were subjected to WGBS on a Novaseq 6000 at Novogene (USA). A Bismark pipeline (v0.22.1) was adapted to process bisulfite sequencing DNA methylation data (Krueger and Andrews 2011). Briefly, raw reads were subjected to Trimmomatic trimming (v0.38) (Bolger *et al.* 2014) to remove adaptor and pool-quality sequences. Bowtie2 (v2.3.5.1) (Langmead and Salzberg 2012) built in Bismark was used for the alignment and alignments of duplicated reads were removed before methylation calling. The methylation level per cytosine site of all three sequence contexts (CG, CHG, and CHH) were determined, which were used for identifying differentially methylated regions (DMRs) with the DSS R package (v2.34.0) (github.com/haowulab/DSS).

DNA methylation around genes and on repetitive sequences

Genomic regions (gene body) from the translation start site (TSS) to the translation termination site (TTS), which were based on genomic locations of major transcripts, were equally divided into 200 windows. For each gene, the 2 kb in the 5'-upstream region and the 2 kb in the 3'-downstream region were also extracted. The DNA methylation rate in three sequence contexts (CG, CHG, and CHH) on each window of the gene body or each 20 bp in upstream and downstream regions was separately determined to examine the distribution of DNA methylation on and around genes.

DMRs were located in the three regions, 5' upstream 1 kb, gene body, 3' downstream 1 kb. For each region, the independence between changes of DNA methylation, increased or decreased in the callus versus the seedling, and regulation in gene expression, up- or down-regulated in the callus versus the seedling from DE analysis, was examined through χ^2statistical test. Tests were performed for all three methylation types: CG, CHG, and CHH.

DNA methylation rate per 100 bp of repetitive sequences was determined. Annotation of repetitive types was from EDTA and additional 45S rDNA alignment analysis. Paired t-test was performed between the two tissues: callus and seedling.

Tissue network and principal component analyses of A188 tissues

The A188 tissue network was constructed with the R package WGCNA (version 1.66) (Langfelder and Horvath 2008) using expression of 29,222 genes in 33 RNA-Seq datasets from 11 A188 tissue types. WGCNA was performed to cluster A188 tissue samples with the parameters minModuleSize = 6 and soft-thresholding power = 9. The Gephi software (version 0.9.2) (Bastian *et al.* 2009) was used to visualize tissue networks with the module and connectivity information from the WGCNA result. Principal component analysis (PCA) was also performed using the R functions *prcomp* with the expression per gene averaged from three replicates per tissue type.

Gene ontology (GO) enrichment analysis

The enrichment analyses were performed to determine if a certain GO was over-represented in a selected group of genes. The resampling method in GOSeq (Young *et al.* 2010) was employed.

Results

Chromosome-level A188 assembly

Long reads, representing a 90X coverage, were generated from A188 genomic DNA using the Oxford Nanopore sequencing platform. The N50 of read lengths is 23.9 kb, and the longest read is 264.5 kb (**Figure B.3**). Genome assembly, performed using Canu, resulted in 1,830 contigs, comprising approximately 2.2 Gb of total sequences. The N50 of contigs is 5.99 Mb (**Table B.4**).

Read depths for contigs were assessed using Illumina short reads generated independently from seedling and immature ear DNAs to identify potential contamination from organelle genomes

or extraneous microbial DNA. Contigs from organelle genomes or extraneous microbial DNA were expected to have differential read depths between the two tissues. Based on this strategy, contigs identified as the chloroplast or mitochondrion sequences were replaced respectively with the previously complete assemblies of A188 organelle genomes (Clifton *et al.* 2004; Bosacchi *et al.* 2015) and contigs from extraneous contamination were discarded (**Figure B.4**). The remaining contigs were polished using raw Nanopore data and 80X PCR-free Illumina 2x250 paired-end whole genome sequencing reads (**Table B.5**), followed by the scaffolding with 113 A188 Bionano Genome (BNG) optical maps, for which the total length is 2.17 Gb and the N50 is 103.4 Mb. The BNG aided assembly placed 875 contigs into 39 scaffolds, which consist of 2.15 Gb. Chromosome pseudomolecules were then generated using a genetic map constructed from 100 B73xA188 double haploid (DH) lines. The final assembly (A188Ref1) consists of 2.25 Gb, including 10 chromosomal pseudomolecules, a mitochondrial genome, a chloroplast genome, and 986 scaffolds or contigs (**Table 2.1**).

The base accuracy of the A188Ref1 assembly was estimated at approximately 99.82% using the KAD pipeline (He *et al.* 2020). Approximately 96.4% of the potential errors are in transposons or other repetitive sequences. The estimated accuracy of genic sequences was >99.97%. The completeness of the A188 assembly was assessed using the BUSCO software (Simão *et al.* 2015) and found to contain 97.25% (3,189/3,278) of the Liliopsida core gene set, similar to the 97.36% (3,193/3,278) in the B73 reference genome (B73Ref4) (Jiao *et al.* 2017).

Presence of complex repeats and nuclear organelle sequences in A188Ref1

In total, 86.3% of the A188 genome sequence is annotated as repetitive elements. The long terminal repeat (LTR) retrotransposons *Gypsy* and *Copia* were the most prevalent elements,

consisting of 44% and 23.9% of A188Ref1, respectively (**Figure 2.2**, circos plot (Krzywinski *et al.* 2009)). LTR centromere retrotransposon of maize (CRM) were largely co-localized with centromere-specific satellite repeat CentC, both of which were largely syntenic to the B73 centromeres (Jiao *et al.* 2017). Approximately 8.3% of A188Ref1 is annotated as DNA transposable elements (TEs), including helitron and Miniature Inverted-repeat Transposable Elements (MITEs) (**Figure 2.2**). Major knob clusters were found on the long arm of chromosomes 5 (5L), the short arm of chromosome 6 (6S), 7L, and 8L, and major subtelomeric repeats (4-12-1) were clustered on the distal regions of 1S, 3S, 4S, 5S, and 8L (**Figure 2.2**). Through similarity alignments, the 45S and 5S ribosomal DNA (rDNA) clusters on 6S and 2L were identified, respectively (**Figure 2.2**). Knob and rDNA locations were in agreement with previously reported A188 fluorescent in situ hybridization (FISH) data (Kato *et al.* 2004). Most repetitive components were located in regions of low-recombination contexts except the 5S rDNA locus and subtelomeric clusters (**Figure B.5**).

Nuclear mitochondrial DNA (NUMT) and nuclear plastid (chloroplast) DNA (NUPT) were identified at 10 and 21 genomic loci, respectively (**Figure 2.3a, Figure B.6**). The largest nuclear organelle-like sequence (~136 kb) is a NUMT locus on the short arm of chromosome 8, which contains an array of DNA transposons likely inserted subsequent to the NUMT integration. FISH analysis corroborated the chromosome 8 location and confirmed a homolog on the distal end of 10S (**Figure 2.3b,3c**). In summary, the genomic locations of repetitive sequences and nuclear organelle sequences are largely consistent with previous findings by FISH (Lough *et al.* 2008; Roark *et al.* 2010), supporting the large-scale correctness of A188Ref1.

Gene annotation

Annotation of A188Ref1 was performed using the Maker pipeline with evidence from transcripts assembled with A188 long Nanopore direct cDNA sequencing data, A188 RNA-Seq Illumina short reads, and transcripts from other maize lines, as well as protein sequences from closely related plant species. The Maker genome annotation resulted in 40,747 high-confidence gene models with 62,142 transcripts (A188Ref1a1) (**Table 2.1**). BUSCO evaluation showed that 97.8% Liliopsida conserved genes were in A188Ref1a1. Comparison of protein sequences identified 52,971 orthologous pairs between A188 and B73, consisting of 27,273 A188 genes and 27,529 B73 genes. We also identified 178 gene clusters in A188 each of which contains at least three paralogous genes, comprising in total 694 genes. The clusters of genes encoding pectin methylesterase (PME) were identified on an unanchored scaffold c04_002 (two clusters with 25 and 18 genes), on chromosome 4 (one cluster with nine genes), and on chromosome 5 (one cluster with five genes) (**Figure 2.4a**). Gene clusters also include eight clusters of 42 nucleotide-binding leucine-rich repeat (NLR) disease resistance (R) genes (**Figure 2.4a**). One NLR gene cluster on chromosome 10 has 16 genes homologous to the *rp1* gene that confers resistance to common rust (Hulbert 1997) and was associated with Goss's wilt resistance (Hu *et al.* 2018) (**Figure 2.4b**). Most paralogous clusters were not located in regions with high recombination (**Figure 2.4c**). Exceptions include the *rp1* locus, which has a high level of haplotype instability through frequent recombination among *rp1* paralogs (Bennetzen *et al.* 1988; Sun *et al.* 2001; Smith *et al.* 2004). Divergent *rp1* haplotypes were observed between A188 and B73 that contains 11 *rp1* homologs at the syntenic locus (**Figure 2.4b**).

We identified 2,259 paralogous gene pairs, of which one member was located in a high-recombination chromosomal compartment and the other in a low-recombination compartment (Methods). Comparison of DNA methylation of A188 seedlings found that, on average, both CG and CHG methylation levels, where H represents A, C, or T, were higher near and within low-recombination paralogous genes as compared to high-recombination genes. No obvious differences were observed in CHH methylation (**Figure 2.4d-f**). Comparison of gene expression between members of the paralogous pairs using seedling RNA-Seq data showed most paralogs had similar expression levels and no expression bias to either high- or low-recombination genes was observed for those paralogs that did exhibit differential expression (**Figure B.7**). The result indicated that the genomic context of genes is a driver for a certain epigenomic modification but not a major driver for gene expression.

High-level structural variation between A188 and B73

Structural variation between the A188 and B73 genomes was identified through comparisons of whole genome assemblies of both genomes using SyRI software (**Supplemental_file_2.1**) (Goel *et al.* 2019) and through the analysis of whole genome Illumina sequencing reads with CGRD (Comparative Genomic Read Depth) that is based on quantitative comparison of depths of short reads (**Supplemental_file_2.2, 2.3**) (Peng *et al.* 2019). SyRI revealed ~1.1 Gb of syntenic regions, 2,302 translocations, as well as 4,083 duplications in B73 and 2,333 duplications in A188 using a minimum cutoff of 10 kb for each translocation or duplication event (**Table B.6**). In addition, SyRI identified 441.9 Mb of B73 and 543.8 Mb of A188 DNA sequences that were not aligned with the other respective genome. Further filtering with CGRD that compared read depths between the two genomes, revealing 381.3 Mb of B73-

specific sequences and 409.2 Mb of A188-specific sequences that represent presence and absence variance (PAV) or highly divergent sequences (HDS). These PAV/HDS regions contain 6,728 genes in B73 and 7,301 genes in A188 (**Supplemental_file_2.4, 2.5**). Gene ontology enrichment analysis indicated that genes related to endopeptidase inhibitor activity and extracellular activities are enriched in both PAV/HDS gene sets (**Figure B.8**).

Seventeen large inversions of 0.5 Mb or greater were identified between the two genomes (**Figure 2.5**, **Figure B.9-B.17**, **Table B.7**). Nine of the seventeen inversions are likely errors in B73Ref4 as the newly released B73Ref5 (unpublished version 5 but available from maizeGDB) showed the same orientation as A188Ref1, including the largest inversion region (INV37083 on B73Ref4, 97.8-103.9 Mb on chromosome 4). FISH analysis of A188 and B73 corroborated the absence of inversion INV37083 (**Figure B.18**). Recombination and pairwise linkage disequilibrium (LD, R^2) values among single nucleotide polymorphisms (SNPs) within each inversion were determined, and out of eight remaining inversion candidates, six have recombination frequencies close to 0 and a high mean LD ranging from 0.56 to 0.79 of all pairs of SNPs that are separated by 0.2-0.3 Mb within an inversion, which are much higher than the genome-wide average LD of 0.2 between SNPs in separated by 0.2 Mb (**Table B.7**). These six inversions exhibiting marked recombination suppression characteristic of inversion (Morgan 1950), therefore, are strongly supported. The six inversions range from 0.7 to 2.1 Mb in size, of which two are located close to the centromere of chromosome 2 and four are on 3L, 4L, 5L, and 9L (**Table B.7**). In total, the six inversion sequences harbor 69 genes in B73 and 75 genes in A188. The syntenic relationships of these genes were largely maintained between inverted sequences in the two genomes (example in **Figure 2.5d**), although the gene sequences are divergent in a high degree from each other (**Figure 2.5b, Figure B.10, B11, B12, and B16**). The divergence of these

inversions indicated that the inversions were not recent events maintained in modern maize populations. Admixture structure analysis showed that both A188 and B73 haplotypes of 3/6 inversions exist in teosinte, the maize wild ancestor (**Figure B.19**), and no clear evidence of the haplotype of the remaining three inversions exist in teosinte (**Figure B.20**). Among landraces, both the A188 and B73 haplotypes of all six inversions could be identified.

CGRD analysis also identified an A188 duplication of a 1.8 Mb region between 8.68 Mb and 10.45 Mb on chromosome 4 of B73Ref4 (**Figure 2.5c**). In A188, a portion of the duplication was found in the unanchored scaffold c04_002 while most of the remaining duplicated sequences can be found in chromosome 4 (**Figure 2.5e**). The duplication region overlapped with the *Gametophyte factor1* (*Ga1*) locus conferring unilateral cross-incompatibility (UCI) (Zhang *et al.* 2018). The underlying causal gene of B73, Zm00001d048936, encodes a PME, which is a wildtype allele. We designed a PME DNA probe that is from the duplication and repeatedly matches 35 loci in B73Ref4 and 78 loci in both the region on chromosome 4 and the scaffold c04_002 on A188Ref1. FISH using this probe resulted in strong hybridization signals on A188 chromosome 4S and weak signals on B73 chromosome 4S, indicating that the duplication occurred locally on 4S (**Figure 2.5f**). The B73 Zm00001d048936 gene has no additional homologous copies in B73Ref4 but five homologous sequences can be identified on the duplicated sequence of A188Ref1, including the syntenic gene Zm00056a022745 that is identical to Zm00001d048936. Collectively, the result documented the complexity and the potential dynamic of the *Ga1* locus of maize.

Associating structural variation with phenotypic variation

The CGRD result indicated that A188 had many more copies (A188plus) at a region from 155.23 to 155.24 Mb of chromosome 9 in B73Ref4 (**Figure B.16**). This region includes the *carotenoid cleavage dioxygenase 1* (*ccd1*) gene catalyzing the cleavage of carotenoids to apocarotenoid products, which is located at the White cap locus (*Wc1*) conditioning kernel colors (Vogel *et al.* 2008). SyRI analysis supported a duplication of this region but failed to find a number of copies in A188. SyRI analysis also indicated the duplicated region is embedded in A188-specific sequences (**Figure 2.6a**). Comparison of A188Ref1 with an A188 BNG optical map aligned to the duplication region indicated the incomplete assembly of the region. Previously, tandem repeats of an ~27 kb sequence at the *Wc1* locus were discovered (Tan *et al.* 2017). Each repeat exhibits four discernible sites that can be detected via Bionano analysis, referred to as Type A repeat. Analysis of A188 sequences revealed a repeat variant containing an additional site, referred to as Type B repeat. Based on the BNG map, the A188 genome contains 13 intact tandem copies of the 27 kb sequence, consisting of 9 copies of Type A and 4 copies of Type B repeats, as well as partial copies of the 27 kb sequence on both ends of the array. Each repeat copy contains a *ccd1* gene, indicating at least 13 copies of *ccd1* in A188 (**Figure 2.6b**), consistent with the A188plus result from the CGRD analysis. Neither intact Type A nor B repeat exists in B73, which, however, does contain a *ccd1* gene.

A188 seeds are white, whereas B73 seeds are yellow (see **Figure 2.1**). Analysis of quantitative trait locus (QTL) of kernel colors of B73xA188 DH lines resulted in two major QTLs on chromosomes 6 and 9, both of which were discovered in a previous genome-wide association study (Romay *et al.* 2013), as well as a weaker peak on chromosome 2 (**Figure 2.6c, Figure B.21**).

Two known genes *y1* and *ccd1* in the major peaks are responsible for kernel colors (**Figure 2.6d**) (Buckner *et al.* 1990; Tan *et al.* 2017). The dominant *Y1* allele conditions yellow kernels (Buckner *et al.* 1990). Several variants exist between the A188 *y1* (Zm00056a032392) and B73 *Y1* (Zm00001d036345) alleles, including one amino acid polymorphism (Ser258Thr) in the coding region (**Figure B.22**) and polymorphisms found in 5' upstream and 3' downstream regions, including a $(CCA)_n$ microsatellite variation in the 5' untranslated region (Phelps *et al.* 1996) (**Figure B.23**). Quantitative reverse transcription PCR (qRT-PCR) reveals higher expression of the *y1* gene in B73 relative to A188 (**Figure 2.6e**). In contrast, the B73 *ccd1* expression was much lower than that of A188, presumably due to the differences in copy number (**Figure 2.6e**). Because higher expression of functional alleles of the *ccd1* and *y1* genes is expected to reduce and increase the accumulation of carotenoids, respectively, the differences in the expression of the *ccd1* and *y1* genes in B73 and A188 likely explains yellow kernels of B73 and white kernels of A188 (**Figure 2.6d,6e**). The expression levels of the alleles in two DH lines with different allele combinations of these two loci were similar (**Figure 2.6e**).

In addition to kernel color, QTL analysis of cob glume color of which A188 is white and B73 is red mapped a single strong peak on the short arm of chromosome 1 (LOD=23.8) (**Figure B.24**). *Pericarp color 1* (*P1*) encoding a Myb transcription factor located in the QTL peak was known to regulate pigment genes (Grotewold *et al.* 1994). The CGRD result indicated that B73 had more copies of the *P1* gene than A188, presenting another structural variation event associated with a phenotypic trait (**Figure B.24**).

Distinct gene expression and hypermethylation in calli relative to seedlings

Transcriptomic data were generated for 11 diverse tissues with three biological replicates each. Both principal component analysis and clustering of these tissue samples based on their genome-wide gene expression showed that the callus from tissue culture were closely related to root, leaf base, embryo, and ear, but distinct from middle leaf, leaf tip, and seedlings (**Figure B.25, Supplemental_file_2.6**). A set of 734 callus featured genes were identified that exhibited at least 2-fold up-regulation in the callus as compared to any other tissues (**Supplemental_file_2.7**). Genes involved in cell wall biosynthesis, defense activity, heme binding, transmembrane transport, and transcription regulation are enriched in these featured genes (**Figure B.25**). For example, a number of NLR and defense-related genes, including *Pathogenesis-related protein 1* (*PR1*) (Zm00056a001451), were activated in the callus. The top six enriched transcription factor families are WOX, AUX/IAA, LBD, AP2, WRKY, and NAC, which included homologs of *Baby boom* (AP2) and *Wuschel2* (Wox) genes relevant to cell division and expansion (**Figure B.26**) (Lowe *et al.* 2016).

The callus and seedling tissues were selected for examination of genome-wide DNA methylation levels. The callus exhibited elevated methylation for all three sequence contexts as compared to the seedling, 89.3% vs 85.2% on CG, 74.5% vs 71.9% on CHG, and 3.2% vs 1.5% on CHH (**Table B.8**). The analysis of CG and CHG methylation over all genes did not find major differences between callus and seedling tissue (**Figure 2.7a, 7b**). However, there were major differences in the level of CHH methylation (**Figure 2.7c**). On average, there were no major changes in the level of CG or CHG methylation over repetitive elements but there was a consistent trend for slightly higher CG methylation callus for most classes of repetitive elements ($p<0.0001$

from paired t-tests, **Figure 2.7d, 7e**). Similarly, CHH methylation was slightly higher for most classes of repetitive elements with the most notable increase observed at MITE elements ($p<0.0001$ from paired t-tests, **Figure 2.7f**).

Differentially methylated regions (DMRs) were identified through comparison of the DNA methylation profiles of callus and seedling. In total, 6,927 CG DMRs, 9,631 CHG DMRs, and 11,275 CHH DMRs were identified (**Table B.9**, **Supplemental_file_2.8, 2.9, 2.10**). Hypermethylation in callus relative to seedling was the predominant type of DMRs for both CG and CHH methylation in both genic and intergenic regions while CHG exhibited roughly equal proportions of hyper and hypomethylation DMRs with more hypermethylation in genic regions and more hypomethylation in intergenic regions (**Figure 2.7g**). The analysis of the distribution of DMRs relative to genes revealed that the CG DMRs were enriched near TSS regions while CHH DMRs tended to be found in regions just upstream or downstream of genes, mirroring CHH island distributions (**Figure 2.7h**) (Gent *et al.* 2013; Li *et al.* 2015b). CHG DMRs exhibited different trends for localization for hyper and hypomethylated DMRs with hypermethylation DNAs enriched at TSS and TTS regions and hypomethylated DMRs enriched in gene bodies (**Figure 2.7h**). The high frequency of some types of DMRs near the TSS led us to assess whether these DMRs may be contributing to differential expression in callus relative to seedling tissue. Genes with DMRs were enriched for being differential expression (DE) in seedling relative to callus compared to genes without DMRs (χ^2=20.9, p-value=4.9e-6). Based on prior studies in maize we expected that gains of CG or CHG methylation near the TSS would be associated with down-regulation of expression while gains of CHH upstream of the promoter might be associated with up-regulation of expression (Li *et al.* 2015a; Sartor *et al.* 2019). We found that the DE genes with hypomethylated or hypermethylated DMRs at most regions exhibited roughly similar numbers of

up- and down-regulated with exception at CG hypomethylation at 5' upstream regions of genes and CHG hypomethylation in the gene body, both of which were associated with up-regulation of gene expression in the callus (**Figure 2.7i, Table B.10**). These results reveal dynamic changes in some types of DNA methylation in callus relative to seedling and a marginal association of DNA methylation with gene expression changes.

Discussion

The A188 genome assembly capitalized on long read technologies, here, Nanopore single molecule reads and long-range optical mapping. The quality of the assembly was enhanced by the strategy of comparing read depths of short read data from independent DNA sources to filter contigs before the scaffolding, which eliminated contamination from DNA sequences of organelle genomes and microorganisms and preserved nuclear-integrated organellar sequences. Our assembly added a new and reference genome to the collection of sequenced maize genomes (Schnable *et al.* 2009; Hirsch *et al.* 2016; Jiao *et al.* 2017; Sun *et al.* 2018; Springer *et al.* 2018; Yang *et al.* 2019; Ou *et al.* 2020; Liu *et al.* 2020; Haberer *et al.* 2020).

A novel strategy for the discovery of genome structural variation based on quantitative comparison of depths of sequencing reads, here named Comparative Genomic Read Depth or CGRD. Detailed characterization of genomic structural variation in complex genomes such as maize is challenging. Comparisons using complete genome sequences based on their alignments would be an ideal method to reveal copy number variation and rearrangements. However, technically, alignment-based methods still suffer from repetitive sequences. More critically, finding structural variation with assembled sequences is subjected to the quality of assemblies. Unfortunately, assemblies of most plant genomes or other large complex genomes are generally

not complete or error-free. For example, here we found that B73Ref4 is missing the topmost region of the short arm sequence of chromosome 6 (**Figure B.13**) and includes multiple assembly inversion errors. CGRD based on comparison of depths of short reads complements the approaches that rely on whole genome alignments, including SyRI (Goel *et al.* 2019). In particular, the CGRD pipeline can detect copy number variation missed by SyRI due to incomplete assembly at structurally complex regions. For example, through CGRD, a 1.8 Mb duplication at the *Ga1* locus and a high-copy tandem duplication of *Wc1* in A188 were identified, both of which did not stand out from SyRI analysis. The two methods are complementary in that CGRD captures unbalanced structural variation due to copy number variation rather than balanced structural variation that SyRI can detect. Therefore, the combination of SyRI and CGRD provides an optimal strategy for discovery of genomic structural variation, which is critical for further characterization of their impacts on gene expression and phenotypes.

In addition to the detection of large duplications and inversions, analysis of structural variation elucidated a repetitive structure of the *ccd1* gene, which, in A188, consists of 13 copies. The high copy number of *ccd1* corresponds to the high expression, which was previously observed and presumably leads to a high activity of the carotenoid cleavage enzyme and enhanced carotenoid degradation (Tan *et al.* 2017). Furthermore, the expression of the A188 *y1* allele is low during seed development, while the *y1* expression in B73 is relatively high (Stelpflug *et al.* 2016). Both alleles were highly expressed in some non-seed tissues such as leaf. The A188 *y1* allele is likely functional, producing a low but perceptible level of carotenoid at certain stages of seed development, as evidenced by pale yellow seeds of recombinant DHs with the A188 *y1* allele and the single copy *ccd1* allele of B73. An additional minor QTL was concordant with QTLs from multiple other B73-derived bi-parental populations (Chandler *et al.* 2013). The candidate gene

zep1 (Zm00001d003513), encoding zeaxanthin epoxidase, was previously identified but functional validation is needed (Owens *et al.* 2019). Based on kernel colors of B73 x A188 DH lines, the three QTL loci are not sufficient to fully determine kernel colors. Analysis with a larger B73xA188 derived population may reveal additional loci influencing kernel colors.

Plant tissue culture from a highly differentiated tissue to the callus involves a process of dedifferentiation to gain pluripotency (Ikeuchi *et al.* 2013). The transition of differentiation status is, physiologically, stressful (Hu *et al.* 2020). Somaclonal variation in plants produced through tissue culture may be the product of DNA damaging stress responses (Lee and Phillips 1988). Transcriptomic data from this study revealed that, in fact, defense response genes were enriched among the callus featured genes that were up-regulated in the callus as compared to any other tissues. Hypermethylation is considered to be a protection mechanism against stresses, which enhance genome stability and safeguard genome integrity (Boyko *et al.* 2007). Our comparison between the callus and the seedling uncovered globally elevated methylation in the callus in all three sequence contexts. Consistently, hypermethylation in the callus relative to the immature embryo was found in a study that used another maize inbred line and methylated DNA immunoprecipitation sequencing (MeDIP-seq) (Liu *et al.* 2017b). In this study, 24 nt small RNA was shown to be positively correlated with DNA methylation. In rice, CG hypermethylation was seen in one- and three-year callus relative to the shoot in the rice mutant *MET1-2*, which encodes a major DNA methyltransferase in maintaining CG methylation. Only CHH hypermethylation was observed in wildtype rice (Hu *et al.* 2020). In our study, the callus versus seedling comparison showed that the A188 *MET1-2* homolog (Zm00056a035610) was ~2x up-regulated in the callus, and *mop1* (Zm00056a013519), a homolog of RNA-dependent RNA polymerase 2 that is involved in the production of 24 nt small RNA (Nobuta *et al.* 2008), was 5-6x up-regulated in the callus,

indicating that the transcriptomic machinery was regulated to enhance global DNA methylation in the callus. In plants regenerated from calli, CG and CHG methylation tended to be lost as compared to non-regenerated plants and many were heritable (Han *et al.* 2018). Heritable hypomethylation in regenerated plants was observed in an earlier maize study (Kaeppler and Phillips 1993). In rice, as compared to non-regenerated plants in rice, pronounced hypomethylation was found in regenerated plants from tissue culture (Stroud *et al.* 2013). The discrepant DNA methylation levels between regenerated plants and calli indicated that most methylation gained from tissue culture is not stable or heritable. Collectively, DNA methylation was elevated during the formation of the callus, likely due to the cellular defense responses. The majority of DNA methylation gained appears to be demethylated during re-differentiation, resulting in hypomethylated regenerated plants.

DH lines have been generated from A188 and B73, a transformation recalcitrant line. Genome-wide genotyping information of DH lines enables the projection of A188 and B73 genome sequences in each DH line. Transformation-enhanced DH lines of B73xA188 crosses, together with the respective reference genome, will provide foundational community resources for insight into maize regeneration.

Conclusions

The genome of a regenerable maize inbred line A188 was assembled with long reads and optical maps, producing a reference-quality genome sequence. Comparison of the A188 genome with the reference B73 genome identified structural variants, including those responsible for phenotypic discrepancies between A188 and B73. Examination of DNA methylation and gene expression with the newly generated A188 reference genome found overly hypermethylation in

the callus as compared with the seedling and the activation of defense genes in the callus, indicative of the defensive state for cellular protection in the embryogenic callus.

Data availability

The A188 genome assembly described in this paper is version JABWIA010000000 at NCBI. The annotation (A188Ref1a1) is available at MaizeGDB.org (https://download.maizegdb.org/Zm-A188-REFERENCE-KSU-1.0/). Raw Nanopore whole genome sequencing data, Illumina whole genome sequencing data, Nanopore cDNA sequencing data, Illumina RNA-seq data, and whole-genome bisulfite are available at NCBI SRA under the project of PRJNA635654. Essential scripts related to the manuscript are available at github.com/liu3zhenlab/A188Ref1. The CGRD pipeline can be downloaded from github.com/liu3zhenlab/CGRD.

References

Alonge M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin, *et al.*, 2019 RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 20: 224.

Armstrong C. L., C. E. Green, and R. L. Phillips, 1991 Development and availability of germplasm with high type II culture formation response. Maize Genet. Coop. News Lett.

Bastian M., S. Heymann, M. Jacomy, and Others, 2009 Gephi: an open source software for exploring and manipulating networks. Icwsm 8: 361–362.

Benjamini Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. 57: 289–300.

Bennetzen J. L., M.-M. Qin, S. Ingels, and A. H. Ellingboe, 1988 Allele-specific and Mutator-associated instability at the Rpl disease-resistance locus of maize. Nature 332: 369–370.

Bolger A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Bosacchi M., C. Gurdon, and P. Maliga, 2015 Plastid genotyping reveals the uniformity of cytoplasmic male sterile-T maize cytoplasms. Plant Physiol. 169: 2129–2137.

Boyko A., P. Kathiria, F. J. Zemp, Y. Yao, I. Pogribny, *et al.*, 2007 Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (virus-induced plant genome instability). Nucleic Acids Res. 35: 1714–1725.

Broman K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

Buckner B., T. L. Kelson, and D. S. Robertson, 1990 Cloning of the y1 locus of maize, a gene involved in the biosynthesis of carotenoids. Plant Cell 2: 867–876.

Bukowski R., X. Guo, Y. Lu, C. Zou, B. He, *et al.*, 2018 Construction of the third-generation Zea mays haplotype map. Gigascience 7: 1-12

Chandler K., A. E. Lipka, B. F. Owens, H. Li, E. S. Buckler, *et al.*, 2013 Genetic analysis of visually scored orange kernel color in maize. Crop Sci. 53: 189–200.

Chu C., C. Wang, C. Sun, C. HSU , K. Yin, *et al.*,1975 Establishment of an efficient medium for another culture of rice through comparative experiments on the nitrogen sources. Sci. Sin. 18: 223–231.

Clifton S. W., P. Minx, C. M.-R. Fauron, M. Gibson, J. O. Allen, *et al.*, 2004 Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol. 136: 3486–3503.

Danecek P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Dobin A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Gent J. I., N. A. Ellis, L. Guo, A. E. Harkess, Y. Yao, *et al.*, 2013 CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res. 23: 628–637.

Gerdes J. T., C. F. Behr, J. G. Coors, and W. F. Tracy, 1993 *Compilation of North American maize breeding germplasm*. Wiley Online Library.

Goel M., H. Sun, W.-B. Jiao, and K. Schneeberger, 2019 SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 20: 277.

Grotewold E., B. J. Drummond, B. Bowen, and T. Peterson, 1994 The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell 76: 543–553.

Haberer G., N. Kamal, E. Bauer, H. Gundlach, I. Fischer, *et al.*, 2020 European maize genomes highlight intraspecies variation in repeat and gene content. Nat. Genet. https://doi.org/10.1038/s41588-020-0671-9

Han Z., P. A. Crisp, S. Stelpflug, S. M. Kaeppler, Q. Li, *et al.*, 2018 Heritable epigenomic changes to the maize Methylome resulting from tissue culture. Genetics 209: 983–995.

He C., G. Lin, H. Wei, H. Tang, F. F. White, *et al.*, 2020 Factorial estimating assembly base errors using k-mer abundance difference (KAD) between short reads and genome assembled sequences. NAR Genomics and Bioinformatics 2.

Hirsch C. N., C. D. Hirsch, A. B. Brohammer, M. J. Bowman, I. Soifer, *et al.*, 2016 Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell 28: 2700–2714.

Holt C., and M. Yandell, 2011 MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.

Hu Y., J. Ren, Z. Peng, A. A. Umana, H. Le, *et al.*, 2018 Analysis of extreme phenotype bulk copy number variation (XP-CNV) identified the association of rp1 with resistance to Goss's wilt of maize. Front. Plant Sci. 9: 110.

Hu L., N. Li, Z. Zhang, X. Meng, Q. Dong, *et al.*, 2020 CG hypomethylation leads to complex changes in DNA methylation and transpositional burst of diverse transposable elements in callus cultures of rice. Plant J. 101: 188–203.

Hufford M. B., X. Xu, J. van Heerwaarden, T. Pyhäjärvi, J.-M. Chia, *et al.*, 2012 Comparative population genomics of maize domestication and improvement. Nat. Genet. 44: 808–811.

Hulbert S. H., 1997 Structure and evolution of the rp1 complex conferring rust resistance in maize. Annu. Rev. Phytopathol. 35: 293–310.

Ikeuchi M., K. Sugimoto, and A. Iwase, 2013 Plant callus: mechanisms of induction and repression. Plant Cell 25: 3159–3173.

Jiao Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, *et al.*, 2017 Improved maize reference genome with single-molecule technologies. Nature 546: 524–527.

Kaeppler S. M., and R. L. Phillips, 1993 Tissue culture-induced DNA methylation variation in maize. Proc. Natl. Acad. Sci. U. S. A. 90: 8773–8776.

Kato A., J. C. Lamb, and J. A. Birchler, 2004 Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proc. Natl. Acad. Sci. U. S. A. 101: 13554–13559.

Kim D., J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, 2019 Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37: 907–915.

Koo D.-H., and J. Jiang, 2009 Super-stretched pachytene chromosomes for fluorescence in situ hybridization mapping and immunodetection of DNA methylation. Plant J. 59: 509–516.

Koo D.-H., F. Han, J. A. Birchler, and J. Jiang, 2011 Distinct DNA methylation patterns associated with active and inactive centromeres of the maize B chromosome. Genome Res. 21: 908–914.

Koo D.-H., H. Zhao, and J. Jiang, 2016 Chromatin-associated transcripts of tandemly repetitive DNA sequences revealed by RNA-FISH. Chromosome Res. 24: 467–480.

Korf I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5: 59.

Kovaka S., A. V. Zimin, G. M. Pertea, R. Razaghi, S. L. Salzberg, *et al.*, 2019 Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20: 278.

Krueger F., and S. R. Andrews, 2011 Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571–1572.

Krzywinski M., J. Schein, I. Birol, J. Connors, R. Gascoyne, *et al.*, 2009 Circos: an information aesthetic for comparative genomics. Genome Res. 19: 1639–1645.

Kurtz S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12.

Langfelder P., and S. Horvath, 2008 WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

Langmead B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Lee M., and R. L. Phillips, 1988 The chromosomal basis of somaclonal variation. Annu. Rev. Plant Physiol. Plant Mol. Biol. 39: 413–437.

Li L., C. J. Stoeckert Jr, and D. S. Roos, 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178–2189.

Li H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26: 589–595.

Li Q., J. Song, P. T. West, G. Zynda, S. R. Eichten, *et al.*, 2015a Examining the causes and consequences of context-specific differential DNA methylation in maize. Plant Physiol. 168: 1262–1274.

Li Q., J. I. Gent, G. Zynda, J. Song, I. Makarevitch, *et al.*, 2015b RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc. Natl. Acad. Sci. U. S. A. 112: 14728–14733.

Li H., 2016 Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32: 2103–2110.

Li H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Liu S., C.-T. Yeh, T. Ji, K. Ying, H. Wu, *et al.*, 2009 Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. 5: e1000733.

Liu S., J. Zheng, P. Migeon, J. Ren, Y. Hu, *et al.*, 2017a Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive Sequences During Maize Domestication and Improvement. Sci. Rep. 7: 42444.

Liu H., L. Ma, X. Yang, L. Zhang, X. Zeng, *et al.*, 2017b Integrative analysis of DNA methylation, mRNAs, and small RNAs during maize embryo dedifferentiation. BMC Plant Biol. 17: 105.

Liu J., A. S. Seetharam, K. Chougule, S. Ou, K. W. Swentowsky, *et al.*, 2020 Gapless assembly of maize chromosomes using long-read technologies. Genome Biol. 21: 121.

Lough A. N., L. M. Roark, A. Kato, T. S. Ream, J. C. Lamb, *et al.*, 2008 Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize. Genetics 178: 47–55.

Lowe K., E. Wu, N. Wang, G. Hoerster, C. Hastings, *et al.*, 2016 Morphogenic regulators baby boom and wuschel Improve monocot transformation. Plant Cell. https://doi.org/10.1105/tpc.16.00124

Martin M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10–12.

McKenna A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Morgan D. T., 1950 A cytogenetic study of inversions in Zea mays. Genetics 35: 153–174.

Nobuta K., C. Lu, R. Shrivastava, M. Pillay, E. De Paoli, *et al.*, 2008 Distinct size distribution of endogeneous siRNAs in maize: Evidence from deep sequencing in the mop1-1 mutant. Proc. Natl. Acad. Sci. U. S. A. 105: 14958–14963.

Ott A., S. Liu, J. C. Schnable, and C. T. E. Yeh, 2017 tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. Nucleic acids.

Ou S., W. Su, Y. Liao, K. Chougule, J. R. A. Agda, *et al.*, 2019 Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20: 275.

Ou S., J. Liu, K. M. Chougule, A. Fungtammasan, A. S. Seetharam, *et al.*, 2020 Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nat. Commun. 11: 2288.

Owens B. F., D. Mathew, C. H. Diepenbrock, T. Tiede, D. Wu, *et al.*, 2019 Genome-wide association study and pathway-level analysis of kernel color in maize. G3: Genes|Genomes|Genetics g3.400040.2019.

Peng Z., E. Oliveira-Garcia, G. Lin, Y. Hu, M. Dalby, *et al.*, 2019 Effector gene reshuffling involves dispensable mini-chromosomes in the wheat blast fungus. PLoS Genet. 15: e1008272.

Phelps T. L., A. E. Hall, and B. Buckner, 1996 Microsatellite repeat variation within the y1 gene of maize and teosinte. J. Hered. 87: 396–399.

Pritchard J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Rakshit S., Z. Rashid, J. C. Sekhar, T. Fatma, and S. Dass, 2010 Callus induction and whole plant regeneration in elite Indian maize (Zea mays L.) inbreds. Plant Cell Tissue Organ Cult. 100: 31–37.

Rhodes C. A., D. A. Pierce, I. J. Mettler, D. Mascarenhas, and J. J. Detmer, 1988 Genetically transformed maize plants from protoplasts. Science 240: 204–207.

Roark L. M., A. Y. Hui, L. Donnelly, J. A. Birchler, and K. J. Newton, 2010 Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. Cytogenet. Genome Res. 129: 17–23.

Romay M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts, *et al.*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14: R55.

Sartor R. C., J. Noshay, N. M. Springer, and S. P. Briggs, 2019 Identification of the expressome by machine learning on omics data. Proc. Natl. Acad. Sci. U. S. A. 116: 18119–18125.

Schnable P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, *et al.*, 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

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

Smit A. F. A., R. Hubley, and P. Green, 2013-2015 RepeatMasker. Open-4.0.

Smith S. M., A. J. Pryor, and S. H. Hulbert, 2004 Allelic and haplotypic diversity at the rp1 rust resistance locus of maize. Genetics 167: 1939–1947.

Springer N. M., S. N. Anderson, C. M. Andorf, K. R. Ahern, F. Bai, *et al.*, 2018 The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat. Genet. 50: 1282–1288.

Stanke M., and S. Waack, 2003 Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2: ii215–25.

Stelpflug S. C., R. S. Sekhon, B. Vaillancourt, C. N. Hirsch, C. Robin Buell, *et al.*, 2016 An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. The Plant Genome 9: lantgenome2015.04.0025.

Steuernagel B., K. Witek, S. G. Krattinger, R. H. Ramirez-Gonzalez, H.-J. Schoonbeek, *et al.*, 2020 The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. Plant Physiol. 183: 468–482.

Stroud H., B. Ding, S. A. Simon, S. Feng, M. Bellizzi, *et al.*, 2013 Plants regenerated from tissue culture contain stable epigenome changes in rice. Elife 2: e00354.

Sun Q., N. C. Collins, M. Ayliffe, S. M. Smith, J. Drake, *et al.*, 2001 Recombination between paralogues at the Rp1 rust resistance locus in maize. Genetics 158: 423–438.

Sun S., Y. Zhou, J. Chen, J. Shi, H. Zhao, *et al.*, 2018 Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat. Genet. https://doi.org/10.1038/s41588-018-0182-0

Tan B.-C., J.-C. Guan, S. Ding, S. Wu, J. W. Saunders, *et al.*, 2017 Structure and origin of the white cap locus and its role in evolution of grain color in maize. Genetics 206: 135–150.

Tang H., J. E. Bowers, X. Wang, R. Ming, M. Alam, *et al.*, 2008 Synteny and collinearity in plant genomes. Science 320: 486–488.

Tang H., X. Zhang, C. Miao, J. Zhang, R. Ming, *et al.*, 2015 ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. 16: 3.

Trapnell C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, *et al.*, 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511–515.

Vega J. M., W. Yu, A. R. Kennon, X. Chen, and Z. J. Zhang, 2008 Improvement of Agrobacterium-mediated transformation in Hi-II maize (Zea mays) using standard binary vectors. Plant Cell Rep. 27: 297–305.

Vogel J. T., B.-C. Tan, D. R. McCarty, and H. J. Klee, 2008 The carotenoid cleavage dioxygenase 1 enzyme has broad substrate specificity, cleaving multiple carotenoids at two different bond positions. J. Biol. Chem. 283: 11364–11373.

Walker B. J., T. Abeel, T. Shea, M. Priest, A. Abouelliel, *et al.*, 2014 Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963.

Wisser R. J., J. M. Kolkman, M. E. Patzoldt, J. B. Holland, J. Yu, *et al.*, 2011 Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. Proc. Natl. Acad. Sci. U. S. A. 108: 7339–7344.

Wu Y., P. R. Bhat, T. J. Close, and S. Lonardi, 2008 Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet. 4: e1000212.

Yang N., J. Liu, Q. Gao, S. Gui, L. Chen, *et al.*, 2019 Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat. Genet. 51: 1052–1059.

Young M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack, 2010 Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11: R14.

Zhang M., Y. Zhang, C. F. Scheuring, C.-C. Wu, J. J. Dong, *et al.*, 2012 Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. Nat. Protoc. 7: 467–478.

Zhang Z., B. Zhang, Z. Chen, D. Zhang, H. Zhang, *et al.*, 2018 A PECTIN METHYLESTERASE gene at the maize Ga1 locus confers male function in unilateral cross-incompatibility. Nat. Commun. 9: 3678.

Figures

Figure 2.1 Phenotypic comparison between A188 and B73

Seed photos of A188 and B73.

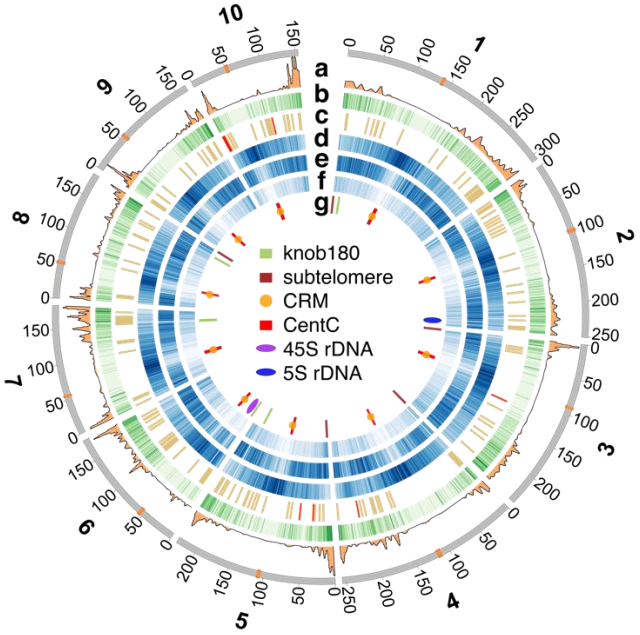

Figure 2.2 Circos plot of A188 genomic features

Features on chromosomes 1 to 10 are: **a**) recombination rate (cM/Mb); **b**) gene density per Mb; **c**) gene clusters; **d**) number of *Gypsy* per Mb; **e**) number of *Copia* per Mb; **f**) number of MITEs per Mb; **g**) high-copy repetitive elements. The central inset is the legend for the track of **g**. Tracks of **b**, **d**, **e**, **f** are intensity-coded. The higher the intensity, the higher frequency each element.

Centromeres are in orange on the outmost chromosome track, on which numbers are coordinates in Mb.

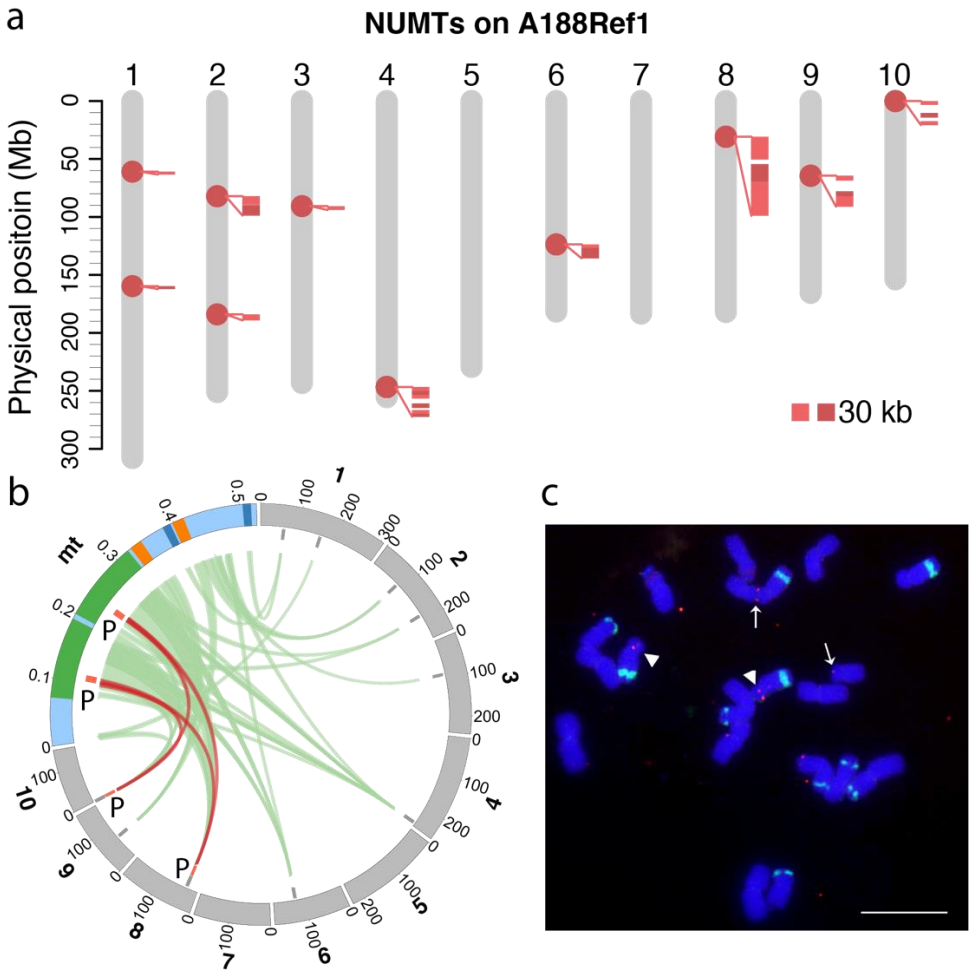

Figure 2.3 NUMT on A188 nuclear genomes

a). NUMT sequence on 10 chromosomes of A188Ref1. Each dot on chromosomes designates a potential NUMT integration. Close-up alignments with the mitochondrion (mt) genome are shown along NUMTs. Each alignment requires at least 5 kb match and 95% identity. **b**). Circos plot of alignments between the mt genome and ten chromosomes. The same color of green, orange, dark blue label duplicated regions in mt. "P" regions match the probe sequence used for FISH. Brown links highlight alignments on chromosomes 8 and 10. Note that the chromosomal scale is different from the mt scale. Numbers on the track are in Mb. **c**). Physical mapping of a mt DNA (mtDNA) and knob repeats on the mitotic metaphase chromosomes of maize A188. The knob repeat probe (green signals) was used to identify the chromosomes. Two FISH sites of mtDNA insertion on the chromosomes were detected: arrowheads, chromosome 8; arrows, chromosome 10. Bar=10 μm.

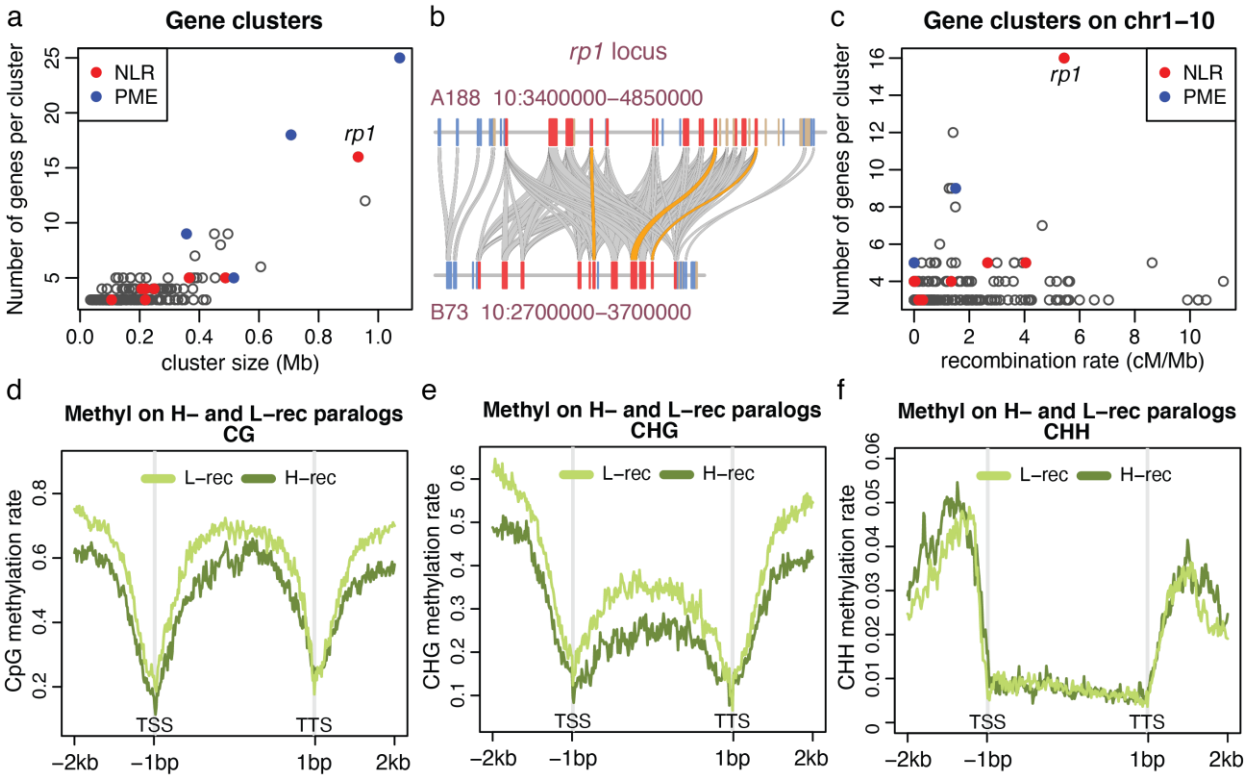

Figure 2.4 Gene clusters and paralogs in low- and high-recombination regions

a). The scatter plot of numbers of genes per cluster versus their cluster size. **b**). Example of an NLR gene (*rp1*) cluster in A188 and their alignments with the B73 *rp1* locus. Each rectangle box represents a gene with blue, tan, and red colors indicating plus, minus orientation, and *rp1* homologous genes. All *rp1* homologs are in the same minus orientation. Gray bands connect orthologs and orange bands highlight the top *rp1* alignments with at least 98.5% identity and a 2,500 bp match. **c**). The scatter plot of numbers of genes per cluster versus the recombination rate estimated 1 Mb around the midpoint of the cluster. All clusters plotted are on 10 chromosomes. **d-f**). Distribution of cytosine methylation in sequence contexts of CG, CHG, and CHH around paralogous genes. An average methylation rate per window across all examined genes from two replicates of seedling samples was determined and plotted versus the window order. A window in the gene body, from translation start site (TSS) to translation termination site (TSS), is 1/200 of the gene body in length. A window outside of the gene body is 20 bp.

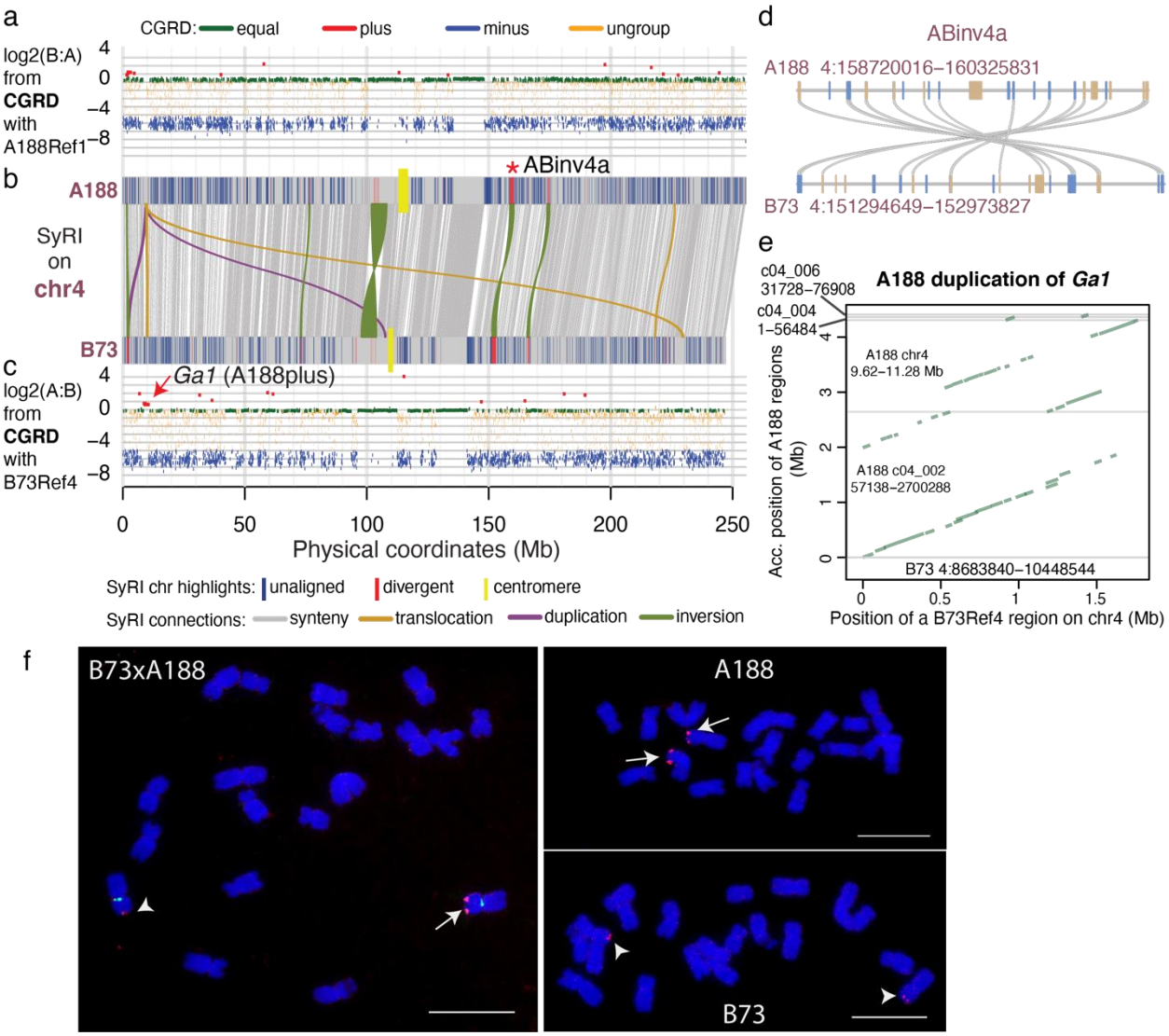

Figure 2.5 Megabase-level duplication and inversion on chromosome 4

a, b, c). SyRI and CGRD results on chromosome 4. **a**). The CGRD result using A188Ref1 as the reference genome. Y-axis represents log2 values of ratios of read depths of B73 to A188, log2(B:A), signifying copy number variation (CNV). Regions with higher and lower sequence depths of B73 versus A188 were B73 plus (red) and B73 minus (blue), respectively. Green and orange represents conserved and ungrouped regions, respectively. **b**). The SyRI result is displayed. Alignments of syntenic blocks larger than 10 kb and alignments of other rearrangements larger than 0.5 Mb are plotted. On each A188 and B73 chromosome, segments that were not aligned to the other genome or highly divergent with the other genome are highlighted. The red * labels a well-evidenced inversion. **c**) The CGRD result using B73Ref4 as the reference genome. The similar color scheme to that in **a**) is used. **d**). Synteny of genes (rectangle blocks) in the well-evidenced inversion (ABinv4a) regions between A188 and B73. blue and tan colors stand for plus

and minus gene orientations. **e**). A dot plot between the 1.8Mb B73 region that was duplicated in A188 and its aligned regions in A188Ref1. **f**). FISH of the PME probe on A188, B73, and F_1 (B73xA188). Cent4 probe (green) that specifically targeted on chromosome 4 centromere was used in F_1 FISH. Arrows and arrow heads point at PME signals of A188 and B73 chromosomes, respectively. Bar=10 μm.

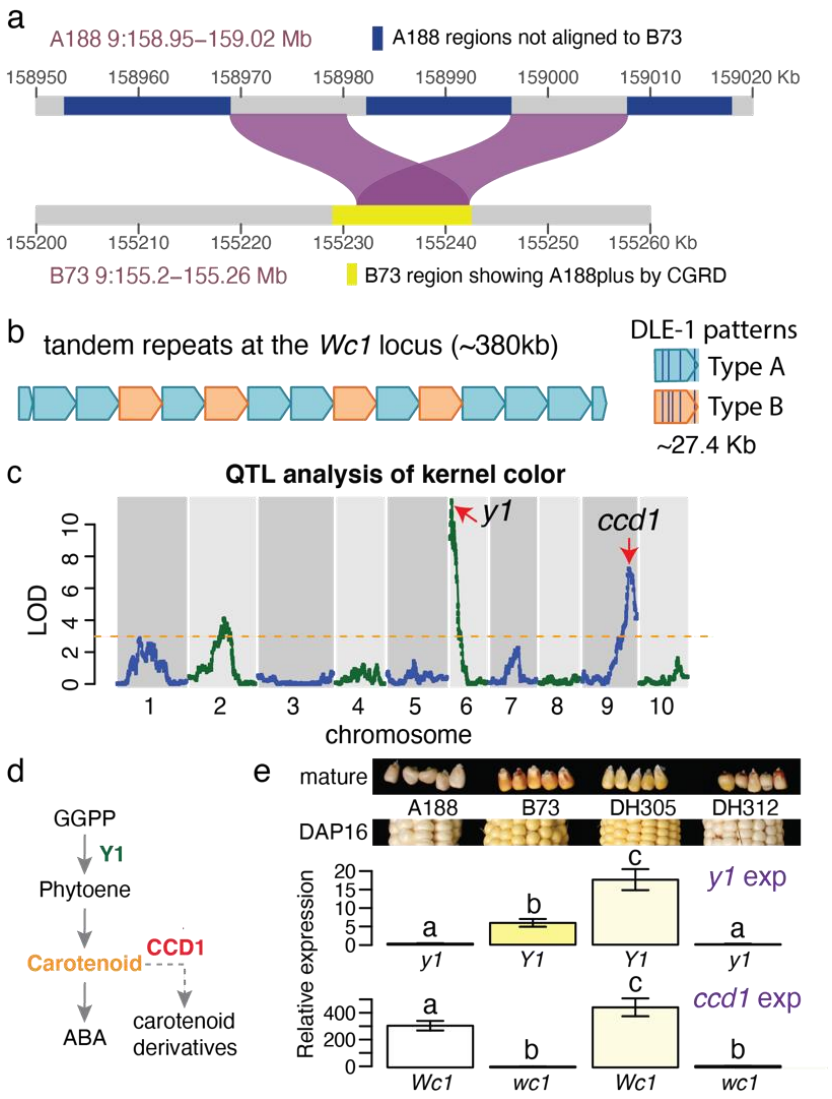

Figure 2.6 Structural variation and genetic analysis of the *Wc1* locus

a). Duplication alignments between A188 and a B73 region identified as A188plus by CGRD. **b**). Tandem repeats of 13 intact copies of 27.4 kb sequences. Two DLE-1 restriction patterns in repeat units: Types A and B were identified. **c**). The QTL result of kernel color using the DH population. Arrows point at locations of known causal genes. **d**). A simplified carotenoid pathway. GGPP stands for geranylgeranyl diphosphate. **e**) Seeds at 16 days after pollination (DAP16) were collected and used for quantifying gene expression (exp) of *y1* and *ccd1*. Three biological

replicates were used. Bars are color-coded based on colors of mature seeds. Error bars represent standard variation. Letters on top of bars are statistical groups determined by Tukey tests. *Y1* (*wc1*) and y1 (*Wc1*) stand for B73 and A188 alleles, respectively. Mature seeds from the same lines show slightly different colors from seeds of DAP16.

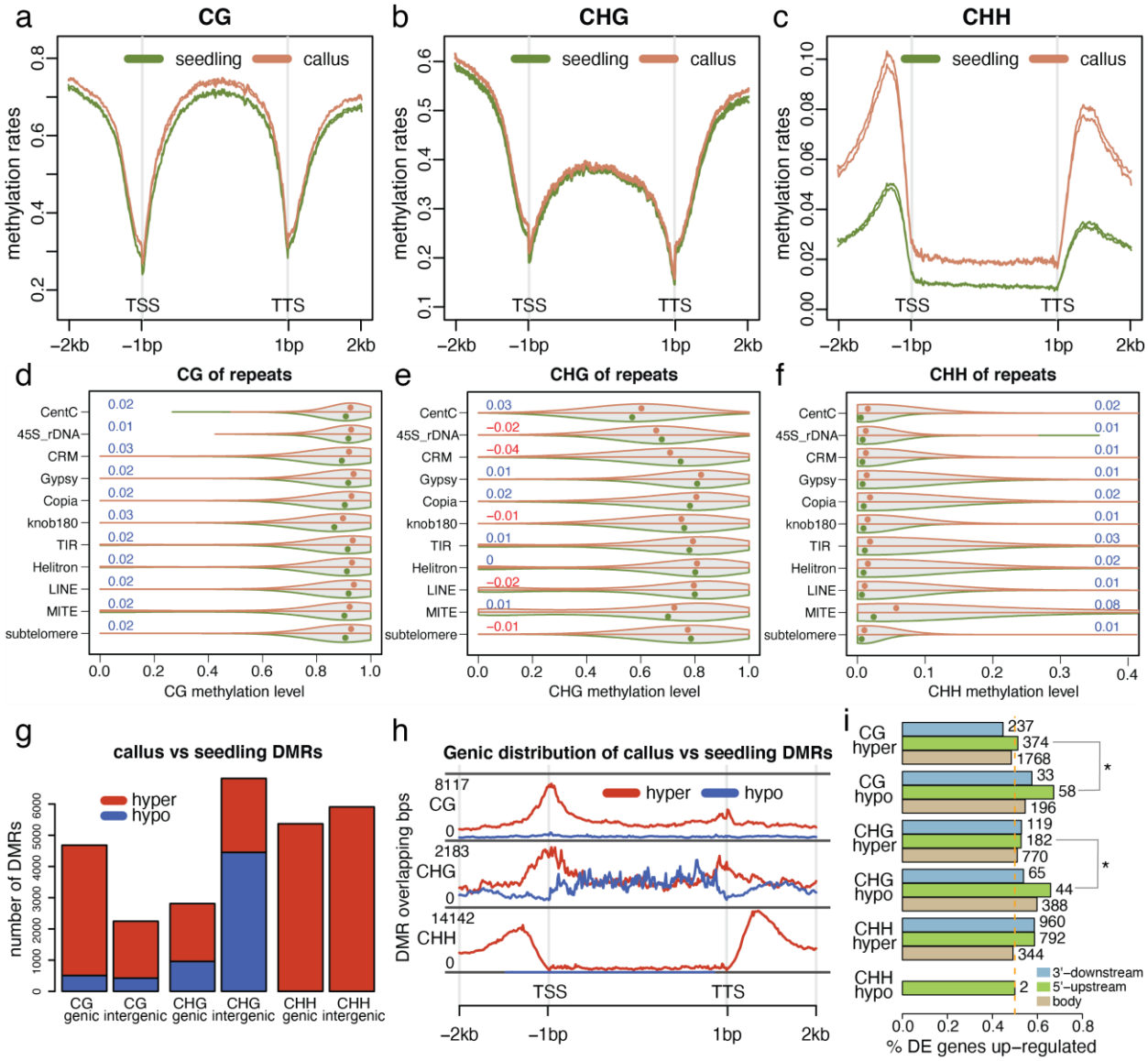

Figure 2.7 DNA methylation in callus and seedling tissues

(**a, b, c**). Distribution of cytosine methylation in three sequence contexts (CG, CHG, and CHH) around genes in two biological replicates of the callus (orange) and two biological replicates of the seedling (green). An average methylation rate per window across all examined genes was determined and plotted versus the window order. A window in the gene body is 1/200 of the gene body. A window outside of the gene body is 20 bp. (**d, e, f**). Violin plots of methylation on repetitive sequences. For each violin plot, the top half is the distribution of methylation in the

callus and the bottom half is the distribution of methylation in the seedling. Each dot represents the median of methylation rates. Numbers stand for the mean methylation differences between the callus and the seedling, which are color coded with blue and red to represent increased methylation and decreased methylation in the callus, respectively. All differences are significant (p-value<0.0001) by paired t-test. (**g**) Barplots of DMRs on genic regions, including 2 kb beyond each of TSS and TTS, and the rest of the genome (intergenic regions). (**h**) Distribution of DMR sequences around genes. The definition of the gene body is the same as described in **a**-**c**. (**i**) Proportions of DE genes up-regulated in hyper DMR and hypo DMR regions (gene body, 1 kb 5' upstream and 1 kb 3' downstream regions). Numbers on top of bars are numbers of DE genes. Stars indicated significances (p<0.05) from χ^2 tests for the independence of the DMR and DE changing directions. In **g**, **h**, and **i,** hyper and hypo stand for increased and decreased methylation in the callus relative to the seedling, respectively.

Table

Table 2.1 Summary of A188Ref1 assembly and annotation

Chromosome	Length (bp)	# genes	# transcripts
1	307,989,483	6,034	9,265
2	251,027,758	4,873	7,384
3	243,219,806	4,313	6,619
4	255,421,021	4,315	6,640
5	229,324,730	4,613	7,194
6	181,596,323	3,412	5,134
7	183,343,242	3,208	4,864
8	182,018,909	3,653	5,472
9	165,494,689	3,082	4,704
10	153,829,095	2,824	4,254
mt	525,405	40	40
pt	140,437	39	41
scaffolds	16,204 -	341	531
sum	2,246,851,412	40,747*	62,142

* filtered from 46,009 gene models produced by Maker

List of supplemental files

Supplemental_file_2.1 SyRI output

Supplemental_file_2.2 CGRD output with B73Ref4 as the reference

Supplemental_file_2.3 CGRD output with A188Ref1 as the reference

Supplemental_file_2.4 A188 genes overlapped with A188-specific sequences

Supplemental_file_2.5 B73 genes overlapped with B73-specific sequences

Supplemental_file_2.6 Normalized read counts from RNA-Seq of 11 tissues

Supplemental_file_2.7 List of callus featured genes

Supplemental_file_2.8 CG DMRs between seedling and callus

Supplemental_file_2.9 CHG DMRs between seedling and callus

Supplemental_file_2.10 CHH DMRs between seedling and callus

Supplemental_file_2.11 The genetic map of BAgm.v02

Chapter 3 - Analysis of Callus Development in Maize via Genetic Mapping and Transcriptional Profiling

Abstract

Plant transformation is generally required for most genome engineering platforms. However, the transformation efficiency is highly dependent on species, individual genotypes, and tissue types. In maize, calli induced from immature embryos are regularly used for transformation. The callus development has been found to be associated with plant regeneration, thereby influencing the transformation efficiency. Of the segregation progeny of a transformation-amenable inbred line A188 and a transformation-recalcitrant inbred line B73, the callus forms into two major types: type I and type II, in which the type II callus grows faster and is the favorable type for regeneration. Here, type I and II calli from the B73xA188 F2 population were analyzed using Genotyping-By-Sequencing (GBS), which identified the quantitative trait loci (QTLs) controlling the callus type on chromosomes 2, 5, 6, 8, and 9. This result was largely supported by the bulk segregant RNA-Seq (BSR-Seq) analysis. Both analyses indicated that only the A188 allele at the chromosome 6 locus positively contributed to the formation of the type II callus. With BSR-Seq, differentially expressed genes (DEGs) between the type II and I F2 calli were identified. In addition, the fast-growth and slow-growth sectors developed from the same A188 immature embryos were separately dissected for the transcriptomic comparison. Both sets of DEGs from the two RNA-Seq comparisons are enriched in the process of cell wall organization, indicating the important role of the cell wall related pathway in callus morphological development. Combination of the five QTLs with the transcriptome analysis identified 39 DEGs located in the broad QTLs interval, providing the candidate genes for plant transformation improvement.

Introduction

Plant transformation is an important process for genome engineering. For both crop improvement and biological research, an efficient transformation in diverse genetic backgrounds is highly beneficial (Que *et al.* 2014; Altpeter *et al.* 2016). However, for many plant species, the transformation efficiency remains low and highly depends on cultivars selected. In maize, even though the transformation frequency has been improved through breeding (Armstrong *et al.* 1991), medium optimization (Duncan *et al.* 1985; Kotchoni *et al.* 2012; Cho *et al.* 2014, 2015), and embryogenesis genes manipulation (Lowe *et al.* 2016), the transformation efficiency across cultivars, or genotypes, varies dramatically, and the underlying genetic basis remains unclear (Que *et al.* 2014; Altpeter *et al.* 2016).

In maize, the embryogenic callus produced by the immature embryo is widely utilized for gene transformation. The callus formation and the regeneration are the major factors influencing the transformation efficiency (Duncan *et al.* 1985; Tomes and Smith 1985; Hodges *et al.* 1986). Of 25 maize inbred lines surveyed in the Hogdes' study (Hodges *et al.* 1986), the calli induced from immature embryos of A188, A634, W117, MS71, and H99 were highly regenerable, while the regeneration frequencies of the calli of B73, H84, and N28 were less than 20%. When the highly regenerable line A188 was crossed to the other 24 inbred lines, the regeneration of the progeny of most of the inbred lines, such as B73, Mo17, H95, Oh43, and VA26, were markedly improved, which suggested the callus embryogenesis is under the genetic control and A188 contains at least some dominant alleles contributing to the regeneration capacity. The genetic elements of the callus embryogenesis were analyzed by QTL mapping (Armstrong *et al.* 1992; Krakowsky *et al.* 2006) and GWAS (Ma *et al.* 2018).

Two distinct types of embryogenic callus, type I and type II, can be initiated from the maize immature embryos. Type I callus was translucent, slow growing, and compact structure mixed with differentiated tissue, while type II callus was highly embryogenic, white or pale yellow, fast growing and friable (Tomes and Smith 1985; D'Halluin *et al.* 1992; Welter *et al.* 1995; Frame *et al.* 2000). Plants are generally regenerated from the type I callus through either the meristem or the somatic embryo, or from the type II callus developed from the somatic embryo (Welter *et al.* 1995). Even though type I and II calli from different genotypes had been used to produce transgenic plants (D'Halluin *et al.* 1992; Ishida *et al.* 1996), type II callus was favorable for gene transformation due to the features of fast growth and high regenerability over years (McCain *et al.* 1988; Frame *et al.* 2000).

Hi-II is a popular line used for maize transformation (Ishida *et al.* 1996; Songstad *et al.* 1996; Que *et al.* 2014), and was generated by the cross of two partial inbred lines, Hi-II A and B, each of which had almost 100% type II callus initiation of immature embryos (Armstrong *et al.* 1991; Ishida *et al.* 1996; Songstad *et al.* 1996). Hi-II A and B were developed from maize inbred lines B73 and A188. The Inbred B73 is an elite line but transformation recalcitrant, while A188 is a highly regenerable line with a poor agronomic performance (Hodges *et al.* 1986). The type II calli can be initiated from both A188 and B73 genotypes but the frequency of type II callus of B73 is much lower (McCain *et al.* 1988). The excellent type II callus initiation of the Hi-II A and B supported the possibility of regeneration improvement through genetic selection.

Plant regeneration can be also improved through molecular manipulation. Two transcription factors BBM and WUS2 have been shown to dramatically improve the embryogenesis in maize and other monocot crops (Lowe *et al.* 2016), indicating the importance of gene regulators for plant regeneration. More studies explored transcriptional regulation during the

callus formation (Shen *et al.* 2012; Salvo *et al.* 2014; Zhang *et al.* 2019; Du *et al.* 2019). Analysis of the callus induction from the maize immature embryo at the early stage revealed that the genes involved in the callus development were enriched in the processes of nutrition uptake, cell wall organization, hormone pathway, stress response, lipid metabolism, signal transduction, oxidation-reduction process, heme binding, and iron ion binding (Shen *et al.* 2012; Salvo *et al.* 2014; Zhang *et al.* 2019; Du *et al.* 2019). In our study, the genetic mapping and transcriptomic profiling were performed to analyze the callus development using both F2 progeny of B73xA188, as well as transcriptomic analysis with fast- and slow-growth callus tissues identified from A188 calli. Our results provide fundamental knowledge for further studies of callus embryogenesis in plants.

Materials and Methods

B73xA188 F2 callus tissues for mapping and RNA-Seq

B73xA188 F1s were grown and self-pollinated to produce F2 ears in the nursery. Immature embryos with length 1.0-1.2 mm were dissected from 13 F2 ears at 11 days after pollination (DAP) and cultured for 3 weeks on N6 medium supplemented with 1.5 mg/L 2,4- dichlorophenoxyacetic acid (2,4-D) at 28˚C in the dark. To map QTLs associated with maize callus type I and II, we separately selected 100 eXtremely type I (XT-I) and II (XT-II) calli from 2,194 F2 calli (13 F2 ears). Each selected callus was cut into two pieces, one was for individual Genotype-By-Sequencing (GBS), and the other one was for bulked segregant RNA-sequencing (BSR-Seq). For BSR-Seq, fifty calli were pooled as one bulk sample, and 4 bulk samples in total were used for RNA-Seq.

A188 callus tissues for RNA-Seq

A188 immature embryos (N=330) were dissected from 4 ears at 11 DAP and cultured on N6 medium for 30 days followed by 5 days sub-culture. Sixty of A188 type II calli with fast and

slow growing sections were sampled. Specifically, on each callus, the fast and slow growing parts were identified and sampled separately. Sections from 20 fast or slow growing calli were separately pooled. In total, 3 fast growing callus bulks and 3 slow growing callus bulks were collected for RNA-Seq.

DNA isolation and GBS sequencing

DNAs of calli was isolated using the DNeasy Plant Mini Kit (Qiagen, USA). In brief, callus samples were disrupted under liquid nitrogen, and then dissolved in buffer AP1 following the manufacturer's instructions. Finally, The DNA was eluted with distilled water and normalized to 15 ng/ul for GBS library preparation.

Hi-II A and B seeds were grown in the greenhouse at 27˚C in the day and 23°C at night with a 16-hour photoperiod. Ten-day-old seedlings were harvested for DNA extraction using the DNeasy Plant Mini Kit (Qiagen, USA). The DNA was dissolved in water and normalized to 15 ng/ul for GBS library preparation.

The GBS protocol was described in chapter 2. In brief, 150 ng DNA of each individual sample was digested with restriction enzymes Bsp1286I (New England Biolabs, USA) at 37˚C for 2 hours followed by the ligation oligos as barcodes using T4 ligase (New England Biolabs, USA) at 16˚C for 1.5 hours. The enzymes in the previous reactions were inactivated at 65˚C for 20 minutes. After that, digestion-ligation products of multiple samples were equally pooled and purified with Qiaquick PCR purification kit (Qiagen, USA) and AMPure XP beads (Beckman Coulter Life Sciences, USA). The purified DNA was input as the template for PCR amplification with the Q5 high fidelity DNA polymerase (New England Biolabs, USA) and the primers matching to Illumina adaptors. The PCR product was purified by using AMPure XP beads (Beckman Coulter

Life Sciences, USA), resulting in a GBS library. The GBS library was sequenced on an Illumina HiseqX 10 platform at Novogene (USA).

RNA extraction and sequencing

XT-I, XT-II, A188 fast and slow growing tissue samples were grounded with mortar and pestle under liquid nitrogen. RNA isolation used the RNeasy Plant Mini Kit (Qiagen, USA) following the manufacturer's instruction. Library preparation and RNA sequencing were performed at Novogene. About 20 million pair reads were generated for each RNA sample on a HiseqX 10 platform.

Genotypes of GBS segment markers

Illumina 150-bp paired raw reads were trimmed using Trimmomatic (version 0.38) (Bolger *et al.* 2014), followed by decoding and fine trimming procedures with custom scripts to assign reads to each individual sample. To increase genotype calling accuracy in the following steps, high-depth (>30x) whole genome sequencing (WGS) data of A188 and B73 were used to call genotypes of the two parental lines (NCBI SRA accession: SRX8420667, SRX8420668, SAMN05578024, SAMN05578025).

After trimming, reads were aligned to the B73 reference genome version 4 (B73Ref4) (Jiao *et al.* 2017) using the BWA aligner (Li and Durbin 2010). Aligned reads were removed if they did not match the following criteria: insert size of 50-800 bp, mapping score greater than 40, the match region greater than 50, the mismatch percentage less than 6%, and the percentage of the unmatched overhang, or the tail, of the read length less than 5. The GATK haplotypecaller (Li and Durbin 2010; McKenna *et al.* 2010; Poplin *et al.* 2018) was used to discover SNPs. SNPs were further filtered and converted to segment (bin) markers by using the R package of Genomap (https://github.com/liu3zhenlab/genomap).

QTL mapping using R/qtl

Genetic position of each segment marker was estimated using a B73XA188 DH genetic map. The R/qtl (Broman *et al.* 2003) function scanone was used to map the QTLs with the standard interval mapping method and the binary model. Two LOD thresholds were used: the LOD value at the 5% significance level from 1000 permutations and the LOD value of 3 (Broman 2001). R/qtl functions plotPXG and fitqtl were used to plot and estimate the QTL effect, respectively.

Genetic mapping using logistic regression

With GBS segment genotyping data of individual XT-I and XT-II, the logistic regression was employed to test the hypothesis that there was no genotype frequency difference between XT-I and XT-II groups. Two approaches were employed to determine significance thresholds. Multiple tests were accounted for with the false discovery rate (FDR) of 1% (Benjamini and Hochberg 1995). A permutation test similar to a standard QTL permutation test was conducted 1000 times to determine the distribution of p-values under the null hypothesis and the significance level of 5% were selected as the p-value cutoff.

BSR-Seq analysis

RNA-Seq raw reads were trimmed using Trimmomatic (version 0.38) (Bolger *et al.* 2014), and then aligned to B73Ref4 with STAR (version 2.7.3a) (Dobin *et al.* 2013). SNPs were discovered using GATK unifiedgenotyper (Li and Durbin 2010; McKenna *et al.* 2010; Poplin *et al.* 2018). Bi-allelic SNPs were selected using GATK with the criteria "AF >= 0.2 && QUAL >= 30.0 && DP >= 100 && DP < 10000". Potential error SNPs based on polymorphic data between B73 and A188 were discarded. The generalized linear model assuming the binomial distribution of two alleles were employed to detect the association between the SNP site and callus types.

Differential expression analysis

Reads trimming and alignment were described in the BSR-Seq method. With read counts per gene resulting from STAR analysis, the statistical test with DESeq2 was performed to identify differentially expressed genes (Love *et al.* 2014). Multiple tests were accounted for with the false discovery rates (FDR) for the XT-II versus XT-I comparison and the fast- versus slow-growth A188 callus comparison (Benjamini and Hochberg 1995).

Gene ontology (GO) enrichment analysis

Differentially expressed genes (DEGs) with the FDR of 10% were used to determine if DEGs are enriched in certain GO terms. The resampling method in GOSeq (Young *et al.* 2010) was employed. The p value cutoff is 0.05.

Identification of candidate genes

Three sets of genes were selected to explore the genes of interest using Venn Diagram in R (https://www.r-graph-gallery.com/14-venn-diagramm.html). The three gene sets include the genes in the QTL intervals and two DEG sets. Genes in the LOD support QTL intervals were identified using BEDtools (Quinlan and Hall 2010). Significant DEGs between type II and I, and between A188 fast- and slow-growth comparison were selected with absolute fold changes greater than 2 and the FDR of 5%.

The function of candidate genes in B73 were obtained from the annotation file deposited on github (https://github.com/liu3zhenlab/collected_data/tree/master/maize_gene_function) and manually checking on MaizeGDB (https://www.maizegdb.org/).

Plotting

Genotypes of markers were plotted using the function plotPXG in R/qtl. All other plots, including genetic mapping of the QTLs, segments genotypes of HI-II, XT-I and II samples, were plotted with custom R scripts.

Results

Genetic mapping of the callus type

Immature embryos dissected from 13 B73xA188 F2 ears were cultured on N6 callus induction media. After three weeks of culture, compact type I and friable type II calli (**Figure 3.1a**) were observed, and 100 most typical type I (referenced to as extremely type I, or XT-I) and 100 most typical type II (referenced to as extreme type II, or XT-II) calli were sampled. Individual calli were subjected to Genotype-By-Sequencing (GBS) and XT-1 and XT-2 calli were separately pooled, 50 calli per bulk, resulting in two bulks of XT-1 and two bulks of XT-II for bulked segregant RNA sequencing (BSR-Seq).

Out of the 200 calli, GBS data of 153 individuals were produced, and 96,703 SNPs were identified. SNP genotypes of each F2 individual were used to infer chromosomal segments harboring multiple SNP markers with the same genotypes. The number of segments indicated the number of discernible recombination events per F2 individual. After filtering F2 individuals with recombination events higher than expected (Ren *et al.* 2020), 60 XT-I and 58 XT-II individuals were retained. With the segment genotypes, 6,369 GBS segment markers were generated for genetic mapping of the callus type (**Figure 3.1b, 3.1d; Figure C.1-C.10; Table 3.1**). A standard interval QTL mapping approach and a method of logistic regression were employed for the QTL mapping. The two approaches identified highly concordant QTL peaks at chromosome 2, 5, 6, 8 and 9, which were designated as CtAB.2.01, CtAB.5.01, CtAB.6.01, CtAB.8.01, and CtAB.9.01, respectively. The LOD supported QTL intervals were estimated (**Table 3.1**). In total, these five QTLs explained 50.9% of the phenotypic variance.

From BSR-Seq of the bulked XT-I and XT-II samples, 68,147 SNPs were identified. Assuming read counts of two alleles of a SNP site had a binomial distribution, the statistical test found 284 associated SNP sites with a divergent allele distribution between the XT-I and XT-II. Most associated SNPs were located on chromosomes 2, 3, 5, 6, 8, and 9, supporting the GBS mapping result with an additional peak on chromosome 3 (**Figure 3.1c; Figure D.3; Table 3.1**).

Among these five QTLs, CtAB.5.01 was the QTL with the strongest signal (**Figure 3.1b-d, 3.2; Table 1**). The homozygous B73 genotype of CtAB.5.01 was highly enriched in XT-II individuals, while the homozygous A188 genotype was enriched in XT-I individuals (**Figure 3.2d,c**). The phenotypic means of the three genotypes indicated that two alleles of QTL CtAB.5.01 functioned additively, and the B73 allele was favorable for the type II callus. In addition to CtAB.5.01, the B73 alleles were favorable alleles for type II callus phenotype at the QTLs CtAB.2.01, CtAB.8.01, and CtAB.9.01 (**Figure 3.3, 3.4; Table 3.1**), whereas the A188 genotype only positively contributed to type II callus at CtAB.6.01 (**Figure 3.3,3.4; Table 3.1**).

The type II callus favorable alleles of three QTLs selected in Hi-II A and B

Hi-II A and B were genotyped by the GBS method and genotypes of chromosomal segments were inferred, which showed that most chromosome regions are in a homozygous status (**Figure 3.5**). Based on genotypes of segments, 22 breakpoints and 54 recombination breakpoints were found in Hi-II A and B, respectively (**Figure 3.5; Table C.1**). The genotypic data also showed that the favorable alleles for type II callus at the three QTLs, CtAB.2.01, CtAB.6.01, and CtAB.8.01, were selected in both Hi-II A and B. However, the type II callus favorable alleles at the major QTL, CtAB.5.01, and one minor QTL, CtAB.9.01 were not selected.

Differential expression of type II and I F2 calli

BSR-Seq were also used to examine differential expression between XT-I and XT-II calli, identifying 1,193 up-regulated and 1,012 down-regulated differentially expressed genes (DEGs). Gene ontology (GO) term enrichment analysis showed that up-regulated genes were enriched in the pathways related to transmembrane components, lipid metabolism, oxidative response, carbohydrate metabolism, DNA binding, and aspartic-type endopeptidase process, while the down-regulated genes were enriched in the pathways involved in DNA binding, metal ion binding, cell wall organization, and aspartyl esterase process (**Figure 3.6**). Further examination found that 66 down-regulated DEGs were associated with cell wall modification, such as pectinesterase, expansin, xyloglucan endotransglucosylase, and beta-galactosidase.

Differential expressed genes of A188 fast- and slow-growth calli

In the tissue culture, the growth rate of callus tissues varies. Fast- and slow-growth calli initiated from single immature embryos of A188 were sampled for RNA-Seq analysis (**Figure 3.7**). In total, 1,287 up-regulated DEGs and 1,926 down-regulated DEGs in fast-growth calli were identified from the comparison. GO enrichment analysis (**Figure 3.8**) indicated that up-regulated DEGs were enriched in the process related to transcriptional regulation, transmembrane transportation, fatty acid biosynthesis, and heme binding, and the down-regulated DEGs were enriched in the process of oxidative response, metal ion binding, DNA binding, heme binding, and cell wall formation. Of the 83 down-regulated DEGs associated with the GO term of cell wall, 60 genes were overlapped with the DEGs from the XT-II and XT-I comparison, which further supported that the cell wall composition plays a role in the growth rate of calli.

Integration of genetic mapping and DEGs

The QTL intervals include 9,273 annotated genes which was about 23.4% of the total annotated genes in B73 version 4 (Jiao *et al.* 2017). From the QTL intervals, we identified 39 DEGs from the two RNA-Seq comparisons with the fold changes in both comparisons greater than 2 and the 5% FDR (**Figure 3.9; Table 3.2**). QTL CtAB.6.01 contains five such genes. All these five genes were down-regulated in the F2 XT-II versus XT-I comparison, and four genes were down-regulated and one was gene up-regulated in the comparison between A188 fast- and slow-growth calli (**Table 3.2**). The four down-regulated genes in both XT-II and fast-growth calli include *wat1* (walls are thin 1, Zm00001d036123) and *end1* (early nodulin homolog1, Zm00001d036125) that are homologs of genes involved in the nodule development, *si1* (silky1, Zm00001d036425) encoding a DNA binding protein involved in the silk development, and the gene Zm00001d036409 encoding an unknown function protein with C2-C2 zinc finger. All these genes were lowly expressed in immature embryos 13 days after pollination (DAP) and highly or medium expressed in the endosperm except *wat1*, which was highly expressed in the pericarp and the endosperm adjacent to scutellum (Doll *et al.* 2020). The cell wall related gene *wat1* encoding an EamA-like transporter was further examined (Figure 3.10). Comparison the A188 and B73 allele sequences found seven SNPs and a 3-bp insertion/deletion (INDEL) in exons, and two large INDELs (382-bp in an intron and 277-bp in the 3' UTR) and a number of SNPs located in other non-coding sequences. However, the allelic expression in XT-II and XT-I maintained a similar ratio based on read counts of each SNP allele in the fifth exon, and no obvious difference was observed in the transcript at the sequence level.

Discussion

Traits related callus development are not easy to quantify and a few studies have been conducted to map transformation related to genomic loci (Armstrong *et al.* 1992; Lowe *et al.* 2006; Krakowsky *et al.* 2006; Salvo *et al.* 2018). In the two studies using A188 or Hi-II, they were used to cross with other transformation-recalcitrant lines and identified chromosomal segments from transformation-amenable parents or segregation distortion after multiple rounds of selection of the highly transformable progeny (Armstrong *et al.* 1992; Lowe *et al.* 2006). Using this atypical QTL mapping strategy, transformation associated loci were found at chromosomes 1, 2, 3, and 9 using an (A188 x B73) x B73 backcrossed population (Armstrong *et al.* 1992), and mapped to chromosomes 1, 2, 3, 6, and 10 using an FBLL x (FBLL x Hi-II) backcross population (Lowe *et al.* 2006). The chromosome 3 QTL was further examined and mapped to an approximately three megabase region (Salvo *et al.* 2018). In our study, we mapped the genomic loci responsible for the callus type to chromosomes 2, 5, 6, 8, and 9. As the previous reports selected highly transformable lines from the backcrossed progeny, the traits under selection are related to callus initiation, callus growth, and regeneration. In our study, the callus type is the trait examined, at least partially explaining the difference in the QTL identification. Type II callus favorable alleles at five QTL we found are found in both A188 and B73 parents, consistent with the finding that their recombinant inbred lines (e.g., Hi-II A and B) developed a higher proportion of type II calli than the parents during tissue culture. Only the QTL on chromosome 6 carries the A188 allele favoriting the type II callus, indicating this QTL is critical for the formation of the type II callus. The QTL on chromosome 6 overlaps with the chromosome 6 QTL detected in Lowe et al. 2006 study, which supports that the chromosome 6 QTL may be correlated with regeneration.

Callus induction is a complex process of cell dedifferentiation. The biological processes involved in early callus induction were revealed through transcriptional analysis (Shen *et al.* 2012; Salvo *et al.* 2014; Zhang *et al.* 2019; Du *et al.* 2019). In our study, the transcription analysis was performed on calli after 21 days and 35 days. In the comparisons between fast growing type II and slow growing type I, and between the fast growing and slow growing A188 type II calli, genes down-regulated in fast growing calli are enriched in the pathway related to the cell wall organization. The cell wall is crucial for plant growth development (Marowa *et al.* 2016), and the differential expressed genes involved in cell wall organization were also identified in the previous transcriptional analysis of callus induction in maize (Shen *et al.* 2012). The down-regulation of genes functioning in cell wall organization could loosen the cell wall and cellular adhesion (Nishikubo *et al.* 2011; Marowa *et al.* 2016; Wormit and Usadel 2018). The difference in the cell wall component and structure between type I and II may result in different water, ion, oxygen, and nutrient uptake ability, thereby affecting the growing rate, transformation efficiency, and regeneration ability.

The DEGs in different callus types are expected to contribute to the phenotypic variation. We found 39 significant DEGs shared by both RNA-Seq comparisons were located in the QTLs intervals. Of the 39 genes, a gene *wat1* close to our QTL is on chromosome 6. Cell elongation was defected in the *Arabidopsis wat1* mutant, which resulted in an abnormal secondary cell wall in the fiber cell and short stem (Ranocha *et al.* 2010). The down-regulated *wat1* gene in our transcription analysis may be involved in the cell wall modification and callus morphology.

Conclusion

In this study, we employed multiple strategies to understand the genetic basis of the formation of the callus type in A188 x B73. Besides genomic changes in the sequence level, the chromatin state may play important roles in the callus response and regeneration. Epigenetic changes during tissue culture in maize and other species were discussed in the previous chapter and literature (Lee and Seo 2018; Han *et al.* 2018). A comprehensive study combining the genetic mapping, transcriptional analysis, and epigenetic analysis are needed to better understand the genetics of callus type trait.

References

Altpeter F., N. M. Springer, L. E. Bartley, A. E. Blechl, T. P. Brutnell, et al., 2016 Advancing Crop Transformation in the Era of Genome Editing. Plant Cell 28: 1510–1520.

Armstrong C. L., C. E. Green, and R. L. Phillips, 1991 Development and availability of germplasm with high Type II culture formation response. Maize Genetics Cooperation Newsletter 65: 92–93.

Armstrong C. L., J. Romero-Severson, and T. K. Hodges, 1992 Improved tissue culture response of an elite maize inbred through backcross breeding, and identification of chromosomal regions important for regeneration by RFLP analysis. Theoretical and Applied Genetics 84-84: 755–762.

Benjamini Y., and Y. Hochberg, 1995 Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. 57: 289–300.

Bolger A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Broman K. W., 2001 Review of statistical methods for QTL mapping in experimental crosses. Lab Anim. 30: 44–52.

Broman K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

Cho M.-J., E. Wu, J. Kwan, M. Yu, J. Banh, et al., 2014 Agrobacterium-mediated high-frequency transformation of an elite commercial maize (Zea mays L.) inbred line. Plant Cell Rep. 33: 1767–1777.

Cho M.-J., J. Banh, M. Yu, J. Kwan, and T. J. Jones, 2015 Improvement of Agrobacterium-mediated transformation frequency in multiple modern elite commercial maize (Zea mays L.) inbreds by media modifications. Plant Cell Tissue Organ Cult. 121: 519–529.

D'Halluin K., E. Bonne, M. Bossut, M. De Beuckeleer, and J. Leemans, 1992 Transgenic maize plants by tissue electroporation. Plant Cell 4: 1495–1505.

Dobin A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, et al., 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Doll N. M., J. Just, V. Brunaud, J. Caïus, A. Grimault, et al., 2020 Transcriptomics at Maize Embryo/Endosperm Interfaces Identifies a Transcriptionally Distinct Endosperm Subdomain Adjacent to the Embryo Scutellum. Plant Cell 32: 833–852.

Du X., T. Fang, Y. Liu, L. Huang, M. Zang, et al., 2019 Transcriptome Profiling Predicts New Genes to Promote Maize Callus Formation and Transformation. Front. Plant Sci. 10: 1633.

Duncan D. R., M. E. Williams, B. E. Zehr, and J. M. Widholm, 1985 The production of callus capable of plant regeneration from immature embryos of numerous Zea mays genotypes. Planta 165: 322–332.

Frame B. R., H. Zhang, S. M. Cocciolone, L. V. Sidorenko, C. R. Dietrich, et al., 2000 Production of transgenic maize from bombarded type II callus: effect of gold particle size and callus morphology on transformation efficiency. In Vitro Cellular & Developmental Biology-Plant 36: 21–29.

Han Z., P. A. Crisp, S. Stelpflug, S. M. Kaeppler, Q. Li, et al., 2018 Heritable Epigenomic Changes to the Maize Methylome Resulting from Tissue Culture. Genetics 209: 983–995.

Hodges T. K., K. K. Kamo, C. W. Imbrie, and M. R. Becwar, 1986 Genotype specificity of somatic embryogenesis and regeneration in maize. Nat. Biotechnol. 4: 219–223.

Ishida Y., H. Saito, S. Ohta, Y. Hiei, T. Komari, et al., 1996 High efficiency transformation of maize (Zea mays L.) mediated by Agrobacterium tumefaciens. Nat. Biotechnol. 14: 745–750.

Jiao Y., P. Peluso, J. Shi, T. Liang, M. C. Stitzer, et al., 2017 Improved maize reference genome with single-molecule technologies. Nature 546: 524–527.

Kotchoni S. O., P. A. Noumavo, A. Adjanohoun, D. P. Russo, J. Dell'Angelo, et al., 2012 A simple and efficient seed-based approach to induce callus production from B73 maize genotype. Am. J. Mol. Biol. 02: 380–385.

Krakowsky M. D., M. Lee, L. Garay, W. Woodman-Clikeman, M. J. Long, et al., 2006 Quantitative trait loci for callus initiation and totipotency in maize (Zea mays L.). Theor. Appl. Genet. 113: 821–830.

Lee K., and P. J. Seo, 2018 Dynamic Epigenetic Changes during Plant Regeneration. Trends Plant Sci. 23: 235–247.

Li H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589–595.

Love M. I., W. Huber, and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15: 550.

Lowe B. A., M. M. Way, J. M. Kumpf, J. Rout, D. Warner, et al., 2006 Marker assisted breeding for transformability in maize. Mol. Breed. 18: 229–239.

Lowe K., E. Wu, N. Wang, G. Hoerster, C. Hastings, et al., 2016 Morphogenic Regulators Baby boom and Wuschel Improve Monocot Transformation. Plant Cell 28: 1998–2015.

Ma L., M. Liu, Y. Yan, C. Qing, X. Zhang, et al., 2018 Genetic Dissection of Maize Embryonic Callus Regenerative Capacity Using Multi-Locus Genome-Wide Association Studies. Front. Plant Sci. 9: 561.

Marowa P., A. Ding, and Y. Kong, 2016 Expansins: roles in plant growth and potential applications in crop improvement. Plant Cell Rep. 35: 949–965.

McCain J. W., K. K. Kamo, and T. K. Hodges, 1988 Characterization of Somatic Embryo Development and Plant Regeneration from Friable Maize Callus Cultures. Bot. Gaz. 149: 16–20.

McKenna A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, et al., 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Nishikubo N., J. Takahashi, A. A. Roos, M. Derba-Maceluch, K. Piens, et al., 2011 Xyloglucan endo-transglycosylase-mediated xyloglucan rearrangements in developing wood of hybrid aspen. Plant Physiol. 155: 399–413.

Poplin R., V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, et al., 2018 Scaling accurate genetic variant discovery to tens of thousands of samples. Cold Spring Harbor Laboratory 201178.

Que Q., S. Elumalai, X. Li, H. Zhong, S. Nalapalli, et al., 2014 Maize transformation technology development for commercial event generation. Front. Plant Sci. 5: 379.

Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Ranocha P., N. Denancé, R. Vanholme, A. Freydier, Y. Martinez, et al., 2010 Walls are thin 1 (WAT1), an Arabidopsis homolog of Medicago truncatula NODULIN21, is a tonoplast-localized protein required for secondary wall formation in fibers. Plant J. 63: 469–483.

Ren W., X. Gong, K. Li, H. Zhang, F. Chen, et al., 2020 Recombination Pattern Characterization via Simulation Using Different Maize Populations. Int. J. Mol. Sci. 21. https://doi.org/10.3390/ijms21062222

Salvo S. A. G. D., C. N. Hirsch, C. R. Buell, S. M. Kaeppler, and H. F. Kaeppler, 2014 Whole transcriptome profiling of maize during early somatic embryogenesis reveals altered expression of stress factors and embryogenesis-related genes. PLoS One 9: e111407.

Salvo S., J. Cook, A. R. Carlson, C. N. Hirsch, S. M. Kaeppler, et al., 2018 Genetic fine-mapping of a quantitative trait locus (QTL) associated with embryogenic tissue culture response and plant regeneration ability in maize (Zea mays L.). Plant Genome 11: 170111.

Shen Y., Z. Jiang, X. Yao, Z. Zhang, H. Lin, et al., 2012 Genome expression profile analysis of the immature maize embryo during dedifferentiation. PLoS One 7: e32237.

Songstad D. D., C. L. Armstrong, W. L. Petersen, B. Hairston, and M. A. W. Hinchee, 1996 Production of transgenic maize plants and progeny by bombardment of hi-II immature embryos. In Vitro Cellular & Developmental Biology - Plant 32: 179–183.

Tomes D. T., and O. S. Smith, 1985 The effect of parental genotype on initiation of embryogenic callus from elite maize (Zea mays L.) germplasm. Theor. Appl. Genet. 70: 505–509.

Welter M. E., D. S. Clayton, M. A. Miller, and J. E Petolino, 1995 Morphotypes of friable embryogenic maize callus. Plant Cell Rep. 14: 725–729.

Wormit A., and B. Usadel, 2018 The Multifaceted Role of Pectin Methylesterase Inhibitors (PMEIs). Int. J. Mol. Sci. 19. https://doi.org/10.3390/ijms19102878

Young M. D., M. J. Wakefield, G. K. Smyth, and A. Oshlack, 2010 Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11: R14.

Zhang X., Y. Wang, Y. Yan, H. Peng, Y. Long, et al., 2019 Transcriptome sequencing analysis of maize embryonic callus during early redifferentiation. BMC Genomics 20: 159.

Figures

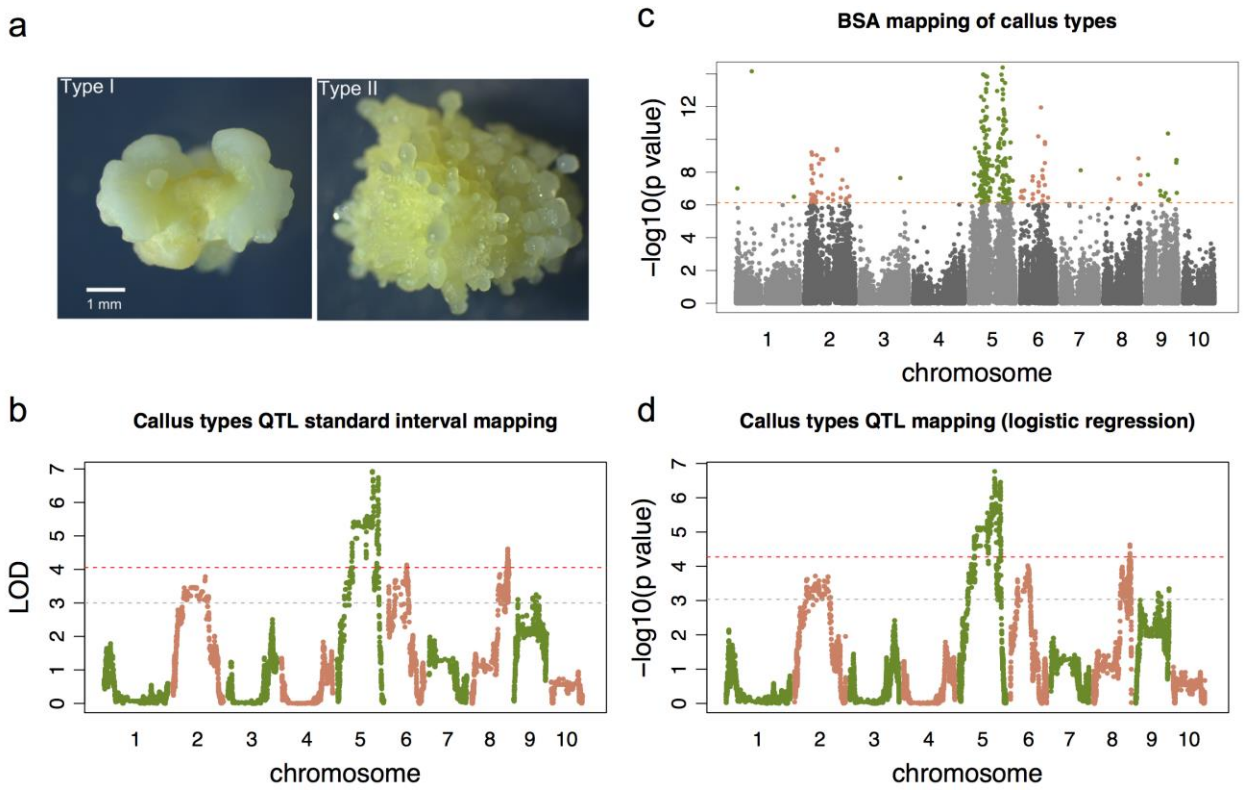

Figure 3.1 Genetic mapping of callus type QTLs

(**a**) Type I and type II calli initiated from B73XA188 F2 immature embryos. (**b,d**) Genetic mapping of callus type with GBS seg markers. **(b**) the QTLs were mapped using the standard interval mapping method (R/qtl) with a binary model. The red dash line indicates the significance threshold defined by the 1000 permutation tests at 5% significance level, and the grey dash line indicates the threshold 3. (**d**) the QTLs were mapped with a logistic regression method. The red dash line indicates the significance threshold defined by the 1000 permutation tests at 5% significance level, and the grey dash line indicates the threshold of FDR=0.01. (**c**) Genetic mapping of callus type with SNPs identified from BSR-Seq. The orange dash line indicates the threshold using the Bonferroni correction at the 5% significance level. The significant SNP markers were colored in olive green or light salmon. In all three mapping plots, x-axes designate the accumulated physical positions of markers.

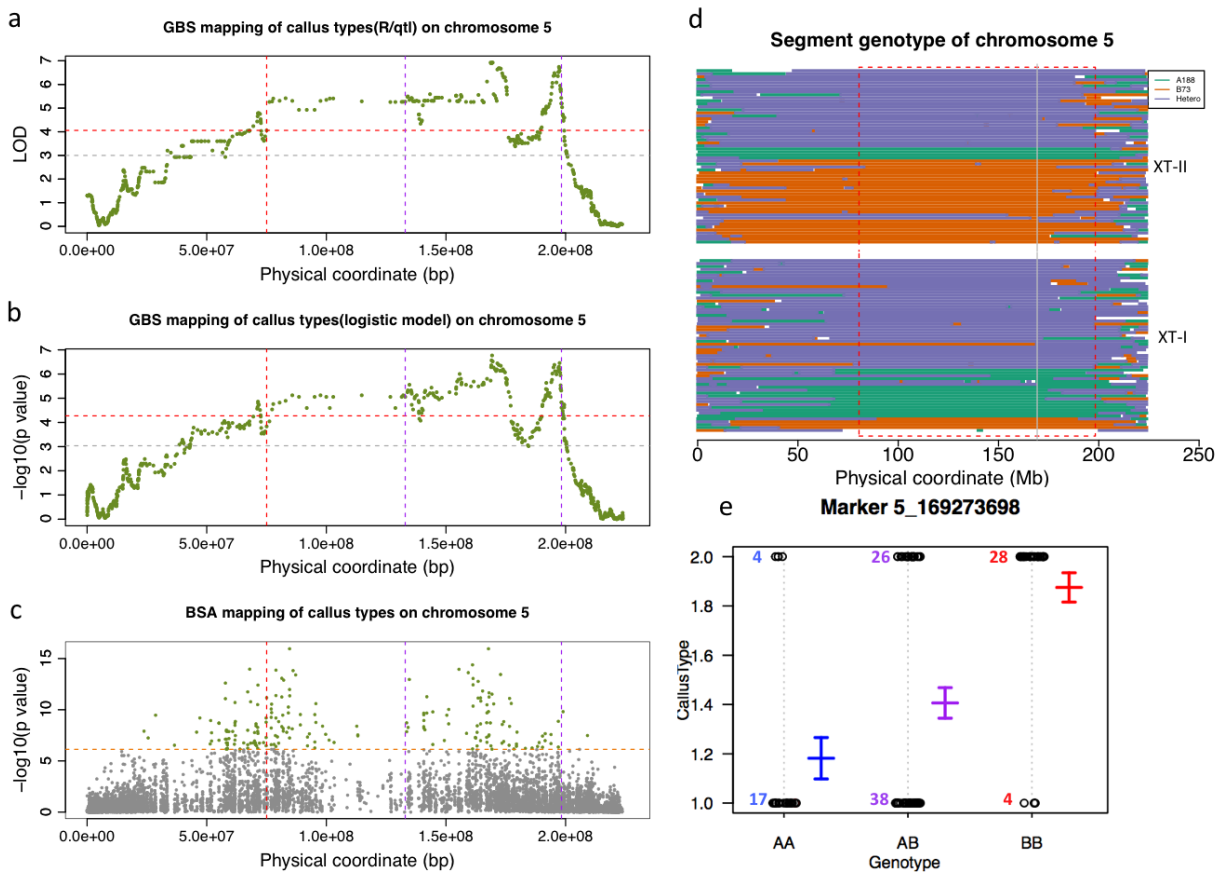

Figure 3.2 Detailed characterization of QTL CtAB.5.01

(**a,b**) Genetic mapping of callus type with GBS segment markers on chromosome 5. (**a**) The result on chromosome 5 using the interval QTL mapping. The red dash line indicates the significance threshold defined by permutation tests at the 5% significance level, and the grey dash line indicates the threshold 3 of LOD. (**b**) The QTL result on chromosome 5 using a logistic regression method. The red dash line indicates the significance threshold with permutation tests at 5% significance level, and the grey dash line indicates the threshold of the 1% FDR. (**c**) Genetic mapping of callus type from BSR-seq. The orange dash line indicates the threshold defined by the Bonferroni correction at the 5% significance level. The significance SNP markers are colored in olive green or light salmon. The vertical purple dash lines (in **a-c**) indicates the LOD support QTL interval, and the red vertical dash line indicates the left flanking of the interval adjusted based on the BSR-seq mapping. (**d**) Genotype of the 118 F2 individuals on chromosome 5. The upper panel contains the genotypes of 58 XT-II individuals, and the bottom panel contains the genotypes of 60 XT-I individuals. Each horizontal line represents the genotype of an individual. The green, orange, purple, white lines stand for a chromosome region with homozygous A188 genotype, homozygous B73 genotype, heterozygous genotype, and missing data, respectively. The red rectangle indicates a QTL interval, and the gray vertical line labels the QTL position mapped by the GBS segment markers. The individuals within a phenotype group (XT-I and XT-II) are ordered based on the genotype of the QTL marker. (**e**) The effect of the QTL marker. The y-axis represents the callus type phenotype, and x-axis

represents different genotypes. Type I and II phenotype is coded as 1 and 2, respectively. The genotype AA is homozygous A188 genotype, AB is heterozygous genotype, and BB is homozygous B73 genotype. The open circle in the plot indicates an F2 individual, which was genotyped by the GBS method. As only two phenotype categories and three genotype groups, the open circles clustered together in different densities. The number of each genotype in each phenotype group is indicated nearby. The phenotype means of each genotype group were plotted in blue, purple and red with standard deviation.

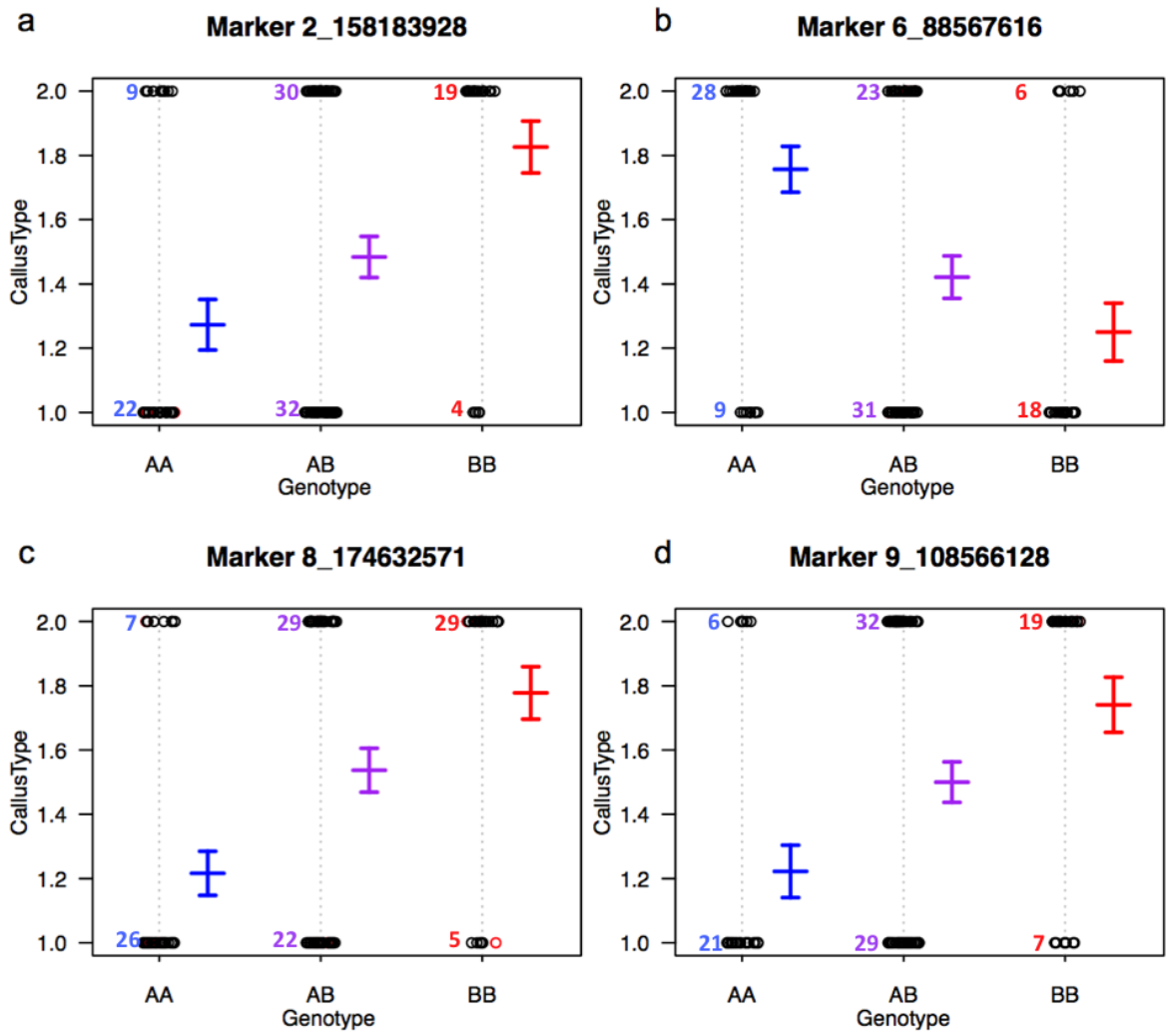

Figure 3.3 Effect of QTL markers on chromosomes 2, 6, 8, and 9

(**a**) Marker effect of QTL CtAB.2.01. (**b**) Marker effect of QTL CtAB.6.01. (**c**) Marker effect of QTL CtAB.8.01. (**d**) Marker effect of QTL CtAB.9.01. The y axis represents the phenotype, and the x axis represents different genotypes. Type I phenotype was coded as 1, and type II was coded as number 2. The genotype AA was homozygous A188 genotype, AB was heterozygous genotype and BB was homozygous B73 genotype. The open circle in the plot indicated an F2

individual, which was genotyped by GBS method. As only two phenotype categories and three genotype groups, the open circles clustered together in different densities. Numbers represent counts of individual samples in each group. The phenotype means of each genotype group were plotted in blue, purple and red with standard deviations.

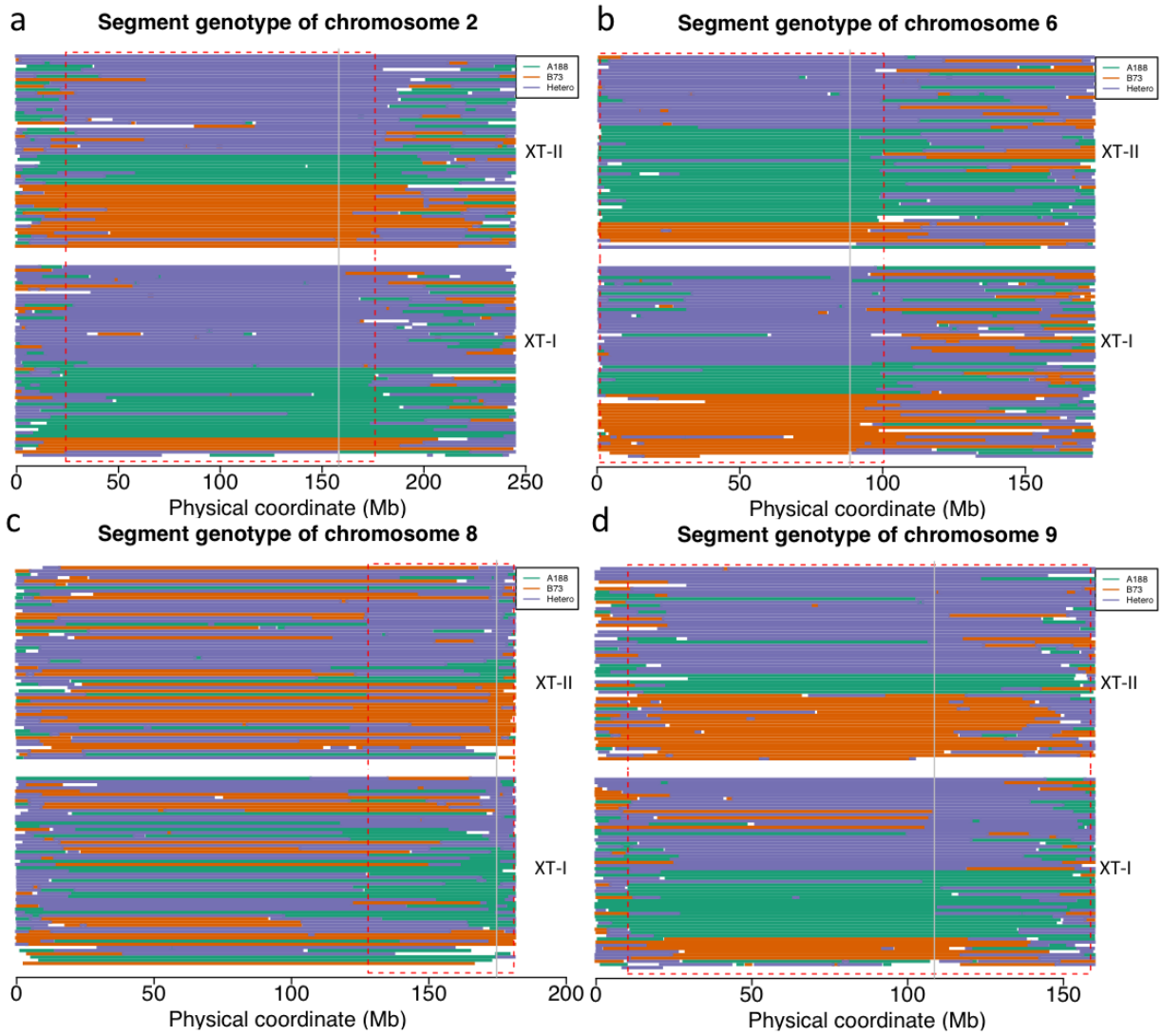

Figure 3.4 Genotype of the 118 F2 individuals on chromosomes 2, 6, 8, and 9

(**a**) Genotype of chromosome 2. (**b**) Genotype of chromosome 6. (**c**) Genotype of chromosome 8. (**d**) Genotype of chromosome 9. The upper panel of each plot includes the genotypes of 58 XT-II individuals, and the bottom panel includes the genotypes of 60 XT-I individuals. Each horizontal line represents the genotype of an individual. The green, orange, purple, white lines indicate the chromosome region (segment) with homozygous A188 genotype, homozygous B73 genotype, heterozygous genotype, and missing data, respectively. The red rectangle indicates the QTL interval, and the gray vertical line labeles at the QTL position mapped by the GBS segment

markers. The individuals within a phenotype group (XT-I and XT-II) were ordered based on the genotype of the QTL marker.

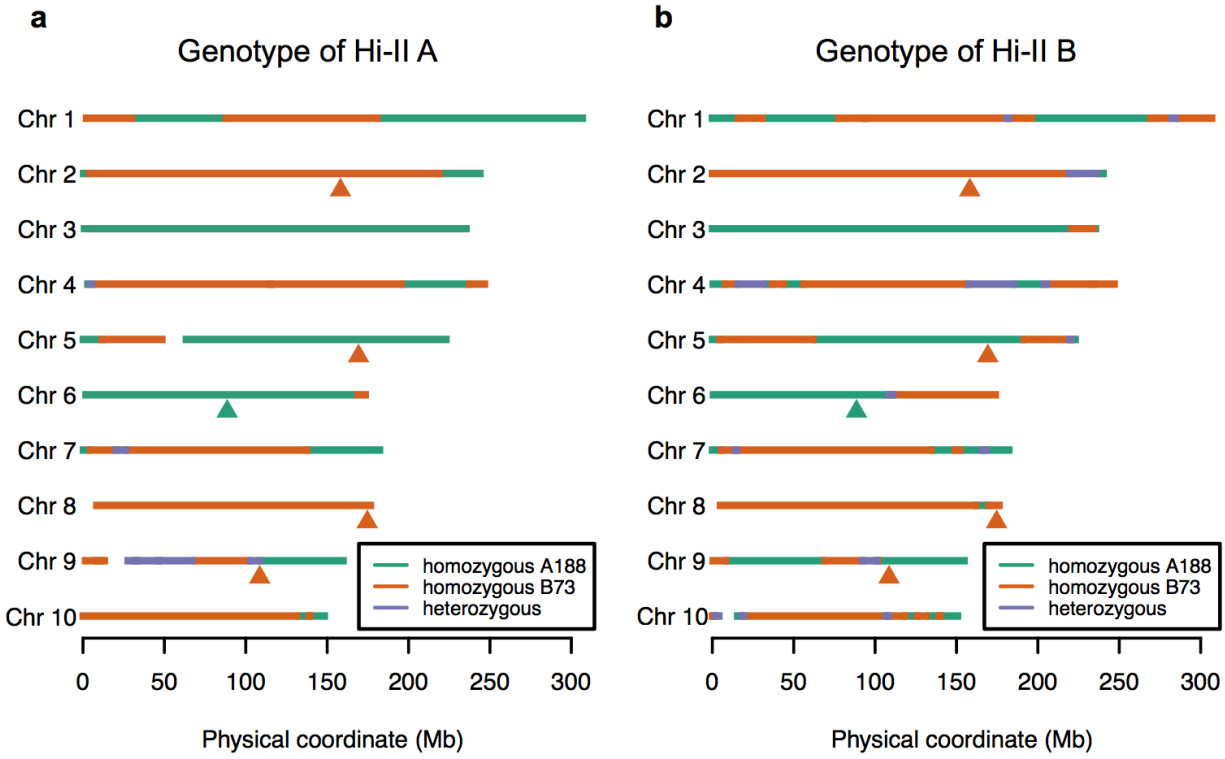

Figure 3.5 Genotypes of Hi-II A and B

(**a**) Genotypes of chromosome segments in Hi-II A. (**b**) Genotype of chromosome segments in Hi-II B. The green, orange, purple, white lines indicate chromosome segment with homozygous A188 genotype, homozygous B73 genotype heterozygous genotype, and missing data, respectively. The triangle under chromosomes labels the QTLs mapped by the GBS method, and the color of the triangle indicates the favorable allele. Green indicates the A188 allele and orange indicates the B73 allele.

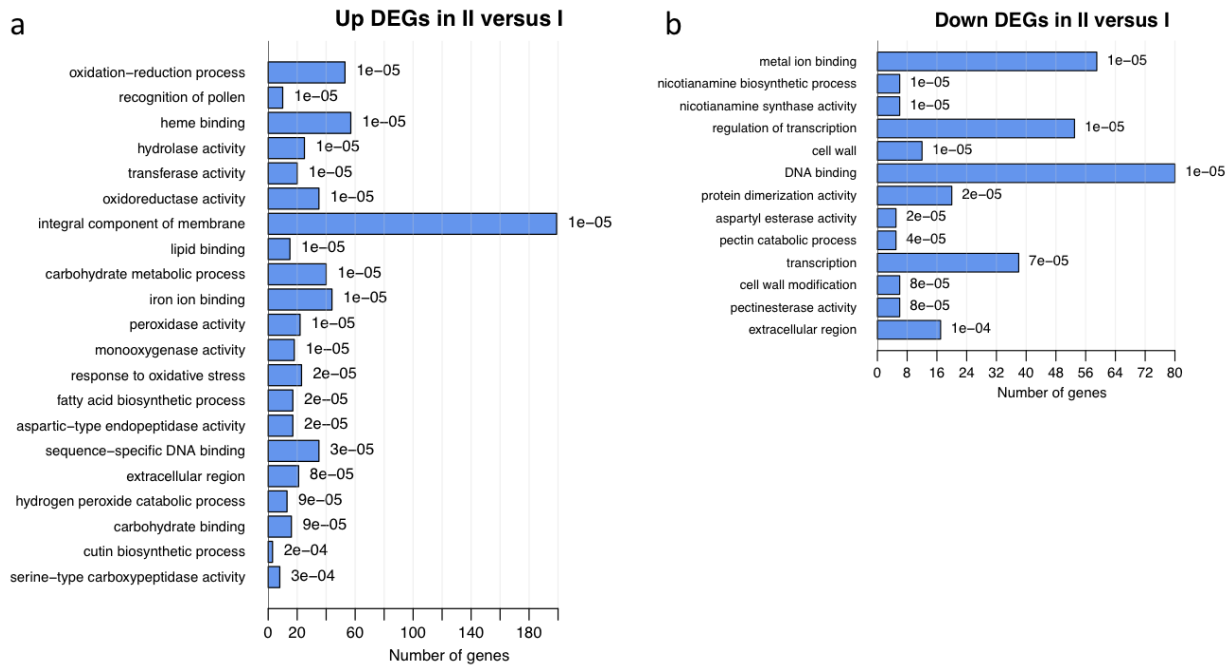

Figure 3.6 GO term analysis of type II and I DEGs

(a) GO term analysis of up-regulated DEGs. (**b**) GO term analysis of down-regulated DEGs. GO terms with a p-value smaller than 0.05 were presented.

Figure 3.7 Fast and slow growing A188 callus

Fast- and slow-growth calli are indicated by ovals and arrows.

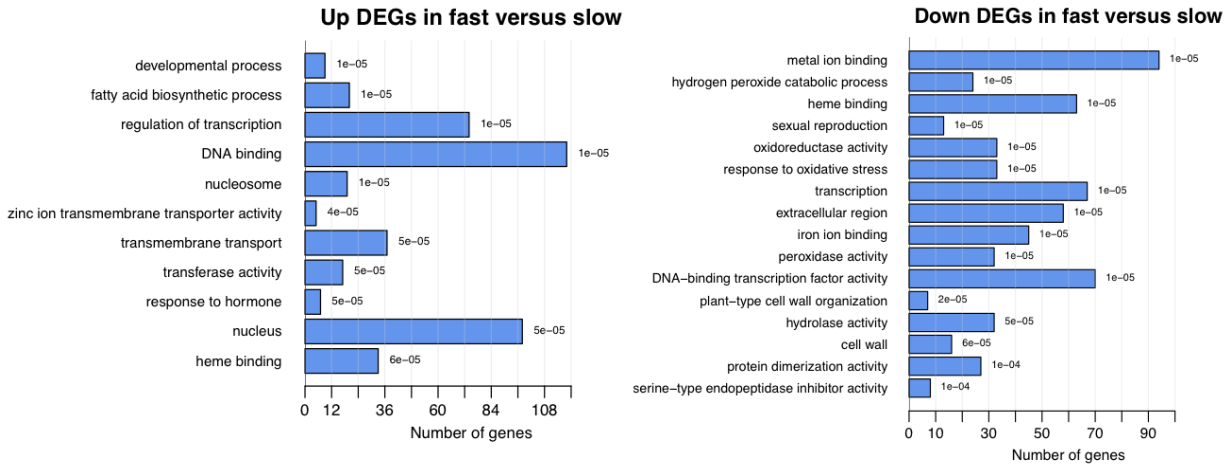

Figure 3.8 GO term analysis of DEGs between fast- and slow-growth A188 calli

(**a**) GO term analysis of up-regulated DEGs. (**b**) GO term analysis of down-regulated DEGs. GO terms with a p-value smaller than 0.05 were presented.

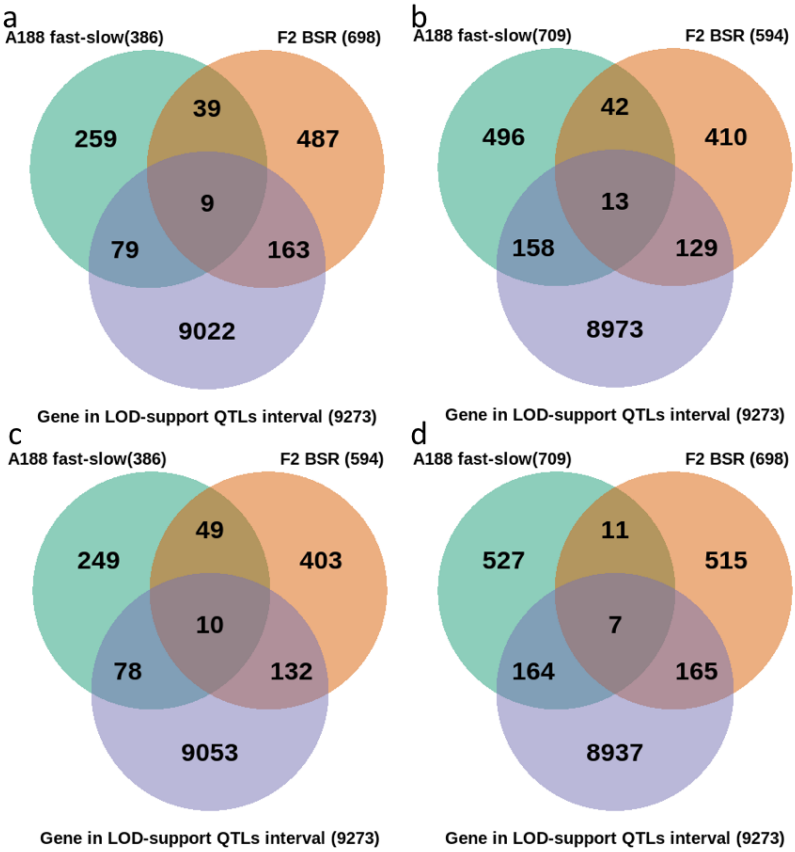

Figure 3.9 Venn diagrams of genes in genetic mapping intervals and DEGs

(**a**) Venn diagram of up-regulated DEGs and genes in the QTLs interval. (**b**) Venn diagram of down-regulated DEGs and genes in the QTLs interval (**c**) Venn diagram of up-regulated DEGs in

the comparison between fast- and slow-growth A188 calli, down-regulated genes in F2 type II and I calli comparison, and the genes in the QTLs interval. (**d**) Venn diagram of down-regulated DEGs in the comparison between fast- and slow-growth A188 calli, up-regulated genes in F2 type II and I calli comparison, and the genes in the QTLs interval.

Figure 3.10 The expression of gene *wat1*

Snapshot of the alignment of RNA-Seq and the alleles of *wat1*. The sample name, DNA sequence name, or the gene annotation are labeled on the left of the alignment or annotation. The gene annotation of *wat1* in B73 Ref4 is indicated in blue rectangle in the bottom track, higher rectangle represents the exon of the gene, and white arrow indicates the transcription orientation. The tracks over the gene annotation represents different alignment files. Reads and coverage are shown in each alignment. The ticks with blue, green, orange, and red indicate the SNPs or sequencing errors. "I" and triangle indicate INDELs. In the DNA sequence Zm00056a032220, the "three reads" in order were the full-length gene annotated in the A188Ref1, and the gaps were the INDELs between A188 and B73 alleles of *wat1*.

Tables

Table 3.1 The QTLs supported by three mapping methods

QTL	Chr	R/qtl mapping		logistic regression		BSR-seq		QTL interval**	
		QTL (bp)	LOD	QTL (bp)	p value	QTL (bp)	p value	Interval start (bp)	Interval end (bp)
CtAB.5.01	5	169273698	6.93	169273698	1.71E-07	84632155	0.00E+00	80458296*	198270432
CtAB.8.01	8	174632571	4.61	174479910	2.35E-05	169073812	1.46E-09	127872514	180929814

CtAB.6.01	6	88567616	4.13	83296319	9.63E-05	99960681	1.12E-12	1006642	100419946
CtAB.2.01	2	158183928	3.79	99580047	1.93E-04	156883753	3.88E-10	24095104	176011606
CtAB.9.01	9	108566128	3.25	159752864	4.50E-04	108540812	4.39E-11	10213832	158723248

**The QTL interval is the LOD support interval, which was estimated in R/qtl.
* QTL CtAB.5.01 was adjusted to include the BSR-seq QTL.

Table 3.2 39 Significant DEGs in the QTL intervals

Gene	Chr.	DEG pattern*	Description
Zm00001d003379	2	up-up	Major latex protein 22
Zm00001d003823	2	up-up	Linoleate 12-epoxygenase
Zm00001d005071	2	up-up	Protein-serine/threonine phosphatase / Serine/threonine specific protein phosphatase
Zm00001d012259	8	up-up	Oligopeptide transporter-related // subfamily not named
Zm00001d012274	8	up-up	Omega-hydroxypalmitate O-feruloyl transferase / O-hydroxycinnamoyltransferase
Zm00001d012535	8	up-up	Family not named // non-specific lipid transfer protein gpi-anchored 2-related
Zm00001d016237	5	up-up	Aquaporin NIP1-1-related
Zm00001d017140	5	up-up	Chlorogenate--glucarate O-hydroxycinnamoyltransferase
Zm00001d017251	5	up-up	Protein CER1-like 1-related
Zm00001d003614	2	down-down	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase
Zm00001d010911	8	down-down	Chitinase / Poly-beta-glucosaminidase
Zm00001d011813	8	down-down	ABA/WDS induced protein (ABA_WDS)
Zm00001d012080	8	down-down	Probable lipid transfer (LTP_2)
Zm00001d015420	5	down-down	Expressed protein
Zm00001d017041	5	down-down	NA
Zm00001d017502	5	down-down	Trehalose-phosphatase / Trehalose 6-phosphate phosphatase
Zm00001d036123	6	down-down	EamA-like transporter family (EamA)
Zm00001d036125	6	down-down	Early nodulin 93 ENOD93 protein (ENOD93)
Zm00001d036409	6	down-down	Zinc-finger of the FCS-type, C2-C2 (zf-FLZ)
Zm00001d036425	6	down-down	MADS box protein // MADS-box family protein

Zm00001d045390	9	down-down	Early nodulin 93 ENOD93 protein (ENOD93)
Zm00001d045391	9	down-down	Early nodulin 93 ENOD93 protein (ENOD93)
Zm00001d003978	2	up-down	NA
Zm00001d003999	2	up-down	Cotton fibre expressed protein (DUF761)
Zm00001d011239	8	up-down	C2 calcium/lipid-binding plant phosphoribosyltransferase-like protein
Zm00001d012477	8	up-down	Laccase / Urishiol oxidase
Zm00001d016632	5	up-down	Serine/threonine-protein kinase rio // subfamily not named
Zm00001d016758	5	up-down	NA
Zm00001d036051	6	up-down	Glucan endo-1,3-beta-D-glucosidase / Laminarinase
Zm00001d045048	9	up-down	Pectinesterase / Pectin methylesterase
Zm00001d045519	9	up-down	ISP4 like protein // subfamily not named
Zm00001d046675	9	up-down	NA
Zm00001d002898	2	down-up	Peroxidase / Lactoperoxidase
Zm00001d002901	2	down-up	Peroxidase / Lactoperoxidase
Zm00001d003432	2	down-up	Aldose 1-epimerase / Mutarotase
Zm00001d004075	2	down-up	Phosphodiesterase I / Phosphodiesterase
Zm00001d004401	2	down-up	Cupin domain (Cupin_2)
Zm00001d004921	2	down-up	O-methyltransferase // subfamily not named
Zm00001d016760	5	down-up	ABA/WDS induced protein (ABA_WDS)

*up-up: Significant DEG was up-regulated in both A188 fast-slow and type II-I comparisons. down-down: both down-regulated. Up-down: up in A188 comparison, down in type II-I comparison; and vice versa.

Chapter 4 - Cloning of the Broadly Effective Wheat Leaf Rust Gene *Lr42* transferred from *Aegilops tauschii*

Abstract

The wheat wild relative *Aegilops tauschii*, a rich resource of genetic variation for wheat improvement, was previously used to transfer the *Lr42* resistance gene into bread wheat. *Lr42* confers resistance at both seedling and adult plant stages to the leaf rust fungus, *Puccinia triticina*. It is broadly effective against all leaf rust races tested to date. *Lr42* has been used extensively in the CIMMYT international wheat breeding program with resulting cultivars deployed in several countries. Using a bulked segregant RNA-Seq (BSR-Seq) mapping strategy, we identified three candidate genes for *Lr42*. A susceptible wheat line transformed with a candidate gene AET1Gv20040300, encoding a nucleotide-binding site leucine-rich repeat (NLR) protein, was sufficient to confer strong resistance to leaf rust. Homologs were found in wheat and wheat relatives, but the *Lr42* resistance allele was found only in one accession of *Ae. tauschii*, suggesting that it is of recent origin. Genetic markers were developed for *Lr42* that identified over one thousand CIMMYT wheat breeding lines that carry the *Lr42* segment, which significantly enhanced resistance to leaf rust at the seedling stage. Cloning of *Lr42* expands the repertoire of cloned rust resistance genes, as well as provides diagnostic DNA markers for wheat improvement. *Lr42* should be deployed in combinations with other leaf rust resistance genes to preserve its utility.

Introduction

Leaf rust, caused by *Puccinia triticina* Erikss., is a prevalent disease limiting wheat production worldwide (Kolmer 1996). Yield reductions in susceptible wheat cultivars typically

range from trace to 30% and may exceed 50%. The yield loss can be mitigated by the introduction of genetic resistance (Huerta-Espino *et al.* 2011). More than 70 leaf rust resistance genes have been characterized and officially named in the Catalogue of Gene Symbols for Wheat (https://shigen.nig.ac.jp/wheat/komugi/genes/symbolClassList.jsp). Six leaf rust resistance genes have been cloned, including four race-specific genes from the nucleotide-binding leucine-rich repeat (NLR) gene family, *Lr10* (Feuillet *et al.* 2003), *Lr21* (Huang *et al.* 2003), *Lr1* (Cloutier *et al.* 2007), and *Lr22*a (Thind *et al.* 2017), as well as two race-nonspecific genes, *Lr34* encoding an ABC transporter (Krattinger *et al.* 2009) and *Lr67* encoding a hexose transporter (Moore *et al.* 2015). Cloned resistance genes may be useful in transgenic multigene cassettes for developing strong and durable resistant varieties to combat fast-evolving fungal pathogens (Wulff and Moscou 2014).

The *Lr42* resistance gene was identified from accession TA2450 in a collection of wheat wild relative *Aegilops tauschii* Coss. (DD, 2n=14), the diploid D-genome donor for hexaploid bread wheat (*Triticum aestivum* L., AABBDD, 2n=42) (Cox *et al.* 1994a). *Lr42* confers resistance to leaf rust at both seedling and adult stages. The *Lr42* gene is effective against all currently reported races of the leaf rust fungus in the US (Kolmer 2019). The *Lr42* resistance locus was introduced to a bread wheat cultivar 'Century' by direct crossing, followed by two backcrosses to Century, and was released in a germplasm line KS91WGRC11 in 1991 (Cox *et al.* 1994b). KS91WGRC11 has been extensively used in CIMMYT wheat breeding programs and is represented as line "*Lr42*" in CIMMYT pedigrees (Basnet *et al.* 2013, 2014). Several cultivars derived from KS91WGRC11 have been released by CIMMYT that have outstanding yield potential. Field studies in Oklahoma showed that near-isogenic lines with *Lr42* introgressions had

a 26% increase in yield and 9% increase in kernel weight, which was attributed to leaf rust resistance (Martin *et al.* 2003).

We undertook the cloning of the *Lr42* gene because of its extensive use in international breeding efforts, broad effectiveness, possible association with yield-enhancing factors, the need for diagnostic markers, and the potential utility of the cloned gene in transgenic cassettes. The *Lr42* gene was previously mapped to the short arm of chromosome 1D (1DS) using hexaploid mapping populations (Cox *et al.* 1994a; Sun *et al.* 2010; Liu *et al.* 2013; Gill *et al.* 2019). We employed BSR-Seq, a bulked segregant RNA sequencing method (Liu *et al.* 2012), to map *Lr42*. To eliminate interference from A-genome or B-genome homoeologous sequences from hexaploid parents, we constructed two diploid mapping populations by crossing the resistant accession with susceptible accessions of *Ae. tauschii*. Another advantage of using diploid parents is that the phenotype of *Lr42* is stronger and easier to distinguish compared to the phenotype in hexaploid wheat. We identified the candidate gene using BSR-Seq and map-based cloning and confirmed the candidate gene as *Lr42* by ectopic expression in a susceptible wheat line.

Materials and Methods

Plant materials

Ae. tauschii accessions used for genetic mapping and haplotype analysis are listed in **Table D.1**. *Ae. tauschii* ssp. *strangulata* accession TA2450 from Caspian Iran is the donor of the *Lr42* gene. Two highly susceptible *Ae. tauschii* accessions TA10132 (also known as AL8/78) and TA2433 were crossed with TA2450 and advanced to $F_{2:3}$, $F_{3:4}$ and $F_{4:5}$ populations by single seed descent for gene mapping and BSR-Seq.

Leaf rust disease phenotyping

The leaf rust disease inoculation procedure followed the protocol developed by Liu *et al*. (Liu *et al.* 2013) except plants were incubated in a growth chamber at 20^{o}C and a 16 hr photoperiod. Briefly, for *Ae. tauschii* leaf rust phenotyping, two-leaf stage seedlings were inoculated with the leaf rust race PNMRJ. The virulence/avirulence phenotype of rust races is given in **Table D.2**. The infection type of plants was scored on the 0 to 4 Stakman scale at 10 days post inoculation (dpi) and confirmed at 14 dpi. The race nomenclature, differential sets, and Stakman infection types were described by Kolmer (1996, 2019). For transgenic wheat seedling leaf rust phenotyping, the race TFBJG was used to inoculate seedlings at the two-leaf stage. TFBJG was used because it defeats *Lr26* in Bobwhite.

Bulked Segregant RNA-Sequencing (BSR-Seq)

Two $F_{2:3}$ populations (population 1: TA2450 x TA2433 and population 2: TA2450 x TA10132) were used for BSR-Seq analysis. Fifteen seeds from each of $F_{2:3}$ families were inoculated and phenotyped. Of 100 TA2450 x TA2433 $F_{2:3}$ families, 27 were homozygous resistant (HR) families of which all individuals were resistant to leaf rust, and 21 were homozygous susceptible (HS) families of which all individuals were susceptible. Equal amounts of leaf tissue were collected from each of 21 HR and 21 HS families, respectively to form two separate pools. Of 101 TA2450 x TA10132 F_3 families evaluated for seedling rust resistance, 36 were HR families and 9 were HS families. We selected 26 HR and all 9 HS families to collect HR and HS tissue pools.

RNA samples were extracted from the seedling tissue using RNeasy Plant Mini Kit (Qiagen, Germany). Sequencing was done on the Illumina HiSeq2000 platform at the Genome Sequencing Facility (GSF) at the Kansas University Medical Center, and ~180 million pairs of 2x101 bp paired-end reads were generated. Raw reads were trimmed using Trimmomatic (version

0.32). Trimmed reads were aligned to the *Ae. tauschii* reference genome (Aet v4.0, GCA_002575655.1)(Luo *et al.* 2017) by GSNAP (version 2018-03-25) (Wu and Nacu 2010) with the parameters of "-B 2 -N 1 -m 6 -i 2 -n 3 -Q". Single nucleotide polymorphisms (SNPs) were discovered using GATK (version 3.3) (McKenna *et al.* 2010) UnifiedGenotyper module with the following parameters: "--heterozygosity 0.005 -stand_call_conf 30.0 -stand_emit_conf 20.0 -glm BOTH -U ALLOW_N_CIGAR_READS -ploidy 2". SNPs was filtered by the GATK module of SelectVariants with the following parameters: --restrictAllelesTo BIALLELIC --selectTypeToInclude SNP --select "AF >= 0.2 && QUAL >= 30.0 && DP >= 200 && DP < 10000". In total, 170,069 SNPs were identified for the population 1 and 74,206 SNPs for the population 2. For each population, a Bayesian-based approach was used to determine the probability of the complete linkage between each SNP and the causal gene (Liu *et al.* 2012).

Fine mapping with KASP markers

Genomic DNA of leaf tissues from *Ae. tauschii* and wheat was extracted as described (Bernardo *et al.* 2020). SNPs having a high probability of the complete linkage with the causal gene were selected to convert to KASP assays. All KASP markers used for fine mapping were listed in **Table D.3**. To confirm the mapping interval, 68 $F_{2:3}$ families from the population 1 used for BSR-Seq were selected for genotyping. The 68 DNAs of pooled tissue samples from 12 $F_{2:3}$ individuals per family were genotyped with KASP markers p12A10, p1A05, and p1A02. As a result, 11 $F_{2:3}$ recombinant families were identified. Analysis of genotypic data together with phenotypic data confirmed that the *Lr42* gene was located between the markers p12A10 and p1A05. To validate this interval, 6 of the 11 recombinant families were selected to genotype individual plants in each family with more KASP markers within the mapping interval.

To narrow down the mapping interval, we used F_4 plants from population 1. We first identified 9 F_4 families that were derived from the resistant $F_{2:3}$ individuals heterozygous for the *Lr42* in the mapping interval. In total, 891 F_4 individuals were phenotyped for rust resistance and genotyped with the markers p12A10 and p1A05, which identified 85 recombinants. Genotyping recombinants with additional markers identified 9 F_4 individuals harboring the recombination between p12A10 and pC24. Further analysis of the F_5 progeny of these 3 F_4 individuals confirmed by the new mapping interval between 8,655,291 bp and 8,830,775 bp on 1DS flanked by the markers pC43 and pC49.

We also analyzed 78 F_2 families of the population 2 and found 4 F_2 families with the recombination between p12A10 and p1A05. Luckily, 1 recombinant between the marker pC43 and the marker pC50 enabled us to locate the gene at a 116 kb interval between 8,655,291 and 8,771,761.

Cloning of full-length coding region of *Lr42* candidate gene

Total RNA was extracted from leaf tissues of resistant (TA2450) and susceptible (TA10132) accessions using TRIzol reagent (Invitrogen, USA) according to the manufacturer's instructions. After removing residual DNA with DNase I (Invitrogen, USA) treatment, 1 µg of total RNA was reverse-transcribed to cDNA using SuperScript® IV First-Strand Synthesis System (Invitrogen, USA) with an oligo$(dT)_{20}$ primer following the manufacturer's instructions. The full-length coding region of the *Lr42* candidate gene was amplified by PCR with the gene-specific primers AET300.2_CDS-F and AET300.2_CDS-R (**Table D.4**). The PCR product was cloned into the pCR-XL-2-TOPO vector (Invitrogen, USA). The inserted fragment in the construct was verified by sequencing using an ABI 3730 DNA analyzer (Applied Biosystems, USA).

Plasmid construction

The full-length of *Lr42* coding regions flanked by a *Bam*HI restriction site was amplified via PCR using primer sets AET300.2_CDS-BamHIF and AET300.2_CDS-BamHIR (**Table D.4**). DNA fragments were ligated into pAHC17 vector (Christensen and Quail 1996) at the *Bam*HI site. The expression constructs contained the full-length *Lr42* coding region under a maize ubiquitin promoter (Ubi-1) and a nopaline synthase terminator (tNOS).

Transgenic plants

Immature embryos were isolated from a spring wheat (*Triticum aestivum* L.) cv. Bobwhite grown in a controlled environment with a 16-h photoperiod, and the day/night temperatures at 20/18°C. The expression constructs and the pAHC20 (Christensen and Quail 1996) containing the *bar* gene were co-bombarded with 1:1 ratio into selected embryogenic calli. Biolistic transformation using a particle inflow gun and following tissue culture protocols were performed as described (Tassy and Barret 2017; Tian *et al.* 2019). Recovered plants in soil were screened for herbicide resistance by brushing a 0.2% v/v Liberty (glufosinate) solution (Bayer CropScience, USA) on leaves. The putative herbicide-resistant plants with an absence of necrosis after 5 days of Liberty application were analyzed by PCR for the presence of the gene of interest (GOI) using primers Ubi-F and Seq2R (**Table D.4**). The transgenic plants with the high expression of GOI were selected for leaf rust bioassays.

***Lr42*-specific GBS tags and identification of *Lr42+* and *Lr42-* CIMMYT wheat lines**

GBS data from both *Ae. tauschii* accessions (Singh *et al.* 2019) and CIMMYT breeding lines (Juliana *et al.* 2019) were used to identify *Lr42*-specific GBS tags from the *Ae. tauschii* donor accession TA2450 after aligning all the GBS tags of TA2450 to the *Ae. tauschii* reference genome (v4.0) (Luo *et al.* 2017). The GBS tags located at the *Lr42* locus (~1 Mb upstream and downstream of the gene) and detected in less than 100 *Ae. tauschii* lines out of all *Ae. tauschii* in the collection

at WGRC were considered to be associated with the *Lr42* segment. From the CIMMYT pedigree, 5,121 CIMMYT lines that were genotyped were involved in the introgression of the *Lr42* segment from TA2450. Given missing data of GBS tags, we expect that each GBS tag that is specifically associated with the *Lr42* segment should be detected in less than 5,000 lines. With that consideration, we obtained 14 *Lr42*-specific GBS tags (**Table D.6**), which were used to identify *Lr42*+ and *Lr42*- wheat lines.

From the CIMMYT pedigrees, 5,121 CIMMYT lines that were GBS genotyped could have the introgression of the *Lr42* segment from TA2450. The wheat lines carrying at least five *Lr42*-specific GBS tags were categorized as *Lr42+*, the lines harboring the *Lr42* segment. The wheat lines with no *Lr42*-specific GBS tags detected but with at least 0.2 million total GBS tags were categorized as *Lr42-*, the lines without the *Lr42* segment. All other lines were not classified.

Phenotypic comparison between *Lr42+* and *Lr42-* CIMMYT lines

Seedling plant responses of CIMMYT lines to leaf rust race MBJ/SP (Lan *et al.* 2014) were obtained using the original disease rating scale of 0-4 and converted to a 0-9 scale for the purpose of quantitative comparison using the conversion formula described in Zhang *et al*. 2014 (Zhang *et al.* 2014). The adult plant scoring was conducted using severity (0-100%, modified Cobb Scale). Seedling leaf rust responses were phenotyped in CIMMYT's greenhouses in El Batan and adult plant leaf rust responses were phenotyped in field trials at two locations, Ciudad Obregon and El Batan, in Mexico. Analysis of variance (ANOVA) and T-test were used to test the association of the *Lr42* marker with seedling and adult plant responses to leaf rust and grain yield measured on a plot-basis in three environments of Obregon: optimum irrigation bed planting, optimum irrigation flat planting and late-sown heat stress bed planting environments. In addition, the same tests were performed on grain yield related traits, such as test weight and thousand kernel weight, evaluated as described by Juliana et al. 2019 (Juliana *et al.* 2019).

Haplotype analysis

DNAs extracted from 35 *Ae. tauschii* accessions in the minicore collection from WGRC were used to survey sequences of *Lr42* haplotypes. *Lr42* alleles/homologs were amplified with the primers Lr42_H1F and Lr42_H1R (**Table D.4**) using Q5® High-Fidelity DNA Polymerase (NEB, USA) with High GC Enhancer. The PCR thermocycling conditions were initial denature at 98˚C for 3 minutes, 33 cycles of 98˚C for 8 seconds, 63˚C for 30 seconds, and 72˚C for 3 minutes, followed by a final extension at 72˚C for 5 minutes. The PCR products were purified by using QIAquick Gel Extraction Kit (Qiagen, Germany) and sequenced by Sanger sequencing in Genewiz (USA). Sequencing reads were *de novo* assembled using Geneious software (version 8.1.7). The command cd-hit-est from the software CD-HIT (4.8.1) was used to cluster *Lr42* allelic homologs with default parameters (Li and Godzik 2006). The allele selected by cd-hit-est to represent each cluster was considered to be the haplotype sequence.

Phylogenetic analysis

Software Geneious (version 8.1.7) was used for multiple alignment and phylogenetic construction. Multiple alignments were performed with the software ClustalW using the default setting. Phylogenetic trees were built with the Juke-Cantor model and the Neighbor-joining method. Trees were exported as Newick formatted flat files that were then uploaded to iTOL for plotting (Letunic and Bork 2019).

Nucleotide diversity

Nucleotide diversity of the 12 *Lr42* alleles was calculated by an R package, PopGenome (Pfeifer *et al.* 2014). The input data for the R package was a ClustalW multiple alignment file in fasta format. Nucleotide diversity was calculated for windows with 50 bp and slided by the step of 10 bp.

Identification of clusters of *Lr42* homologs (*Lr42* cluster)

Clusters of *Lr42* homologs were identified with BLAST (Altschul *et al.* 1990)**.** First, the *Lr42* resistant allele was aligned to the genomes of *Brachypodium* (GCF_000005505.3 Brachypodium distachyon v3.0) (International Brachypodium Initiative 2010), Barley (GCA_901482405.1_ Morex_v1.0) (Mascher *et al.* 2017), *Triticum dicoccoides* wild emmer (GCA_002162155.2 WEW v2.0) (Avni *et al.* 2017), *Triticum turgidum* subsp. durum (GCA_900231445.1 Svevo.v1) (Maccaferri *et al.* 2019), *Ae. tauschii* (Aet v4.0) (Luo *et al.* 2017), and *T. aestivum* cv. CS (iwgsc_refseqv1.0) (International Wheat Genome Sequencing Consortium (IWGSC) *et al.* 2018). Homologs were identified if an alignment had the E-value smaller than 1e-100 and the matched length of the query (*Lr42*) was longer than 1 kb. Second, a chromosome interval smaller than 2 Mb with at least 2 homologs was identified as a *Lr42* cluster. Alignments of the *Lr42* resistant allele and homologs in each cluster were plotted using Circos (Krzywinski *et al.* 2009).

Semi-quantitative RT-PCR

Leaf tissue from *Ae. tauschii* and transgenic wheat were collected, and RNA was extracted using RNeasy Plant Mini Kit (Qiagen, Germany). cDNA was synthesized with Verso cDNA Synthesis Kit (Thermo Scientific, USA). The cDNA input for each sample was normalized by the housekeeping gene *Actin* amplified with primers actin_F1 and actin_R1 (**Table D.4**) for 25 cycles. The *Lr42* resistant and susceptible alleles were amplified with primers Lr42-qRT-F5/R5 and lr42_1F/R (**Table D.4**) for 28 cycles.

Conserved domain and repeats annotation

Protein and DNA sequence was submitted to NCBI for conserved domain search (Marchler-Bauer *et al.* 2017). Leucine-rich repeat (LRR) was searched by a web based LRR search tool with additional manual examination (Bej *et al.* 2014).

Development of diagnostic markers for *Lr42*

Multiple alignment of *Lr42* alleles from the *Ae. tauschii* minicore set identified a unique region (~140bp) in the LRR (*Lr42R*-unique-segment) from the *Lr42* resistant allele. To design diagnostic markers on the *Lr42* gene using this region, we aligned the *Lr42* sequence of the unique region to all *Lr42* homologs in reference genomes of *Ae*. *tauschii*, wild emmer, durum wheat, CS, Barley, and Brachypodium. The top hit was a homolog (1D:7381846-7384626) in CS. The top hit sequence carries two SNPs compared with the *Lr42* unique sequence. Outside this highly similar region, high polymorphisms were found between *Lr42* and the homolog 1D:7381846-7384626. The second best hit sequence has 19 SNPs, confirming that the *Lr42R*-unique-segment is not common in diverse genomes. Based on this finding, for each KASP assay, we design a *Lr42* specific primer on the *Lr42R*-unique-segment, and a primer on a homolog from the cluster of *Lr42* homologs on CS 1D. The common primer paired with them was designed on a conserved region between *Lr42* and the homolog. The primer pair that amplifies the *Lr42* homolog could potentially amplify a paralog in *Ae*. *tauschii* genomes. Therefore, in most populations, the assay is considered to be a dominant marker for detection of the *Lr42* resistant allele.

Results

Genetic mapping pinpointed *Lr42* on 1DS and identified candidate genes

For efficient genetic mapping, we established two diploid mapping populations by crossing the *Lr42* donor *Ae. tauschii* TA2450 with two leaf rust susceptible *Ae. tauschii* accessions, TA2433 (**Figure 4.1a**) and TA10132 (**Figure D.1**). $F_{2:3}$ individuals from each population were phenotyped for leaf rust resistance at the seedling stage. We scored 100 $F_{2:3}$ families of the TA2450 x TA2433 population and identified 27 homozygous resistant and 21 homozygous susceptible $F_{2:3}$ families (**Figure 4.1b**). Leaf tissues of homozygous resistant and susceptible $F_{2:3}$ family seedlings were

separately pooled for BSR-Seq (Liu *et al.* 2012). The BSR-Seq experiment mapped *Lr42* at a locus close to the end of 1DS (**Figure 4.1c**), consistent with the mapping results from the other mapping population TA2450 x TA10132 (**Figure D.1**), and from the previous *Lr42* mapping studies in hexaploid wheat (Sun *et al.* 2010; Liu *et al.* 2013; Basnet *et al.* 2014; Gill *et al.* 2019). The results indicated that the gene we mapped in the diploid populations is the same as the *Lr42* gene transferred to hexaploid wheat. Based on the BSR-Seq results, we identified single nucleotide polymorphisms (SNPs) that were likely located near the *Lr42* gene and converted them to KASP (Kompetitive Allele Specific PCR) markers for genotyping F_3 and F_4 individuals from both mapping populations (**Table D.3**). The *Lr42* mapping interval was narrowed down to approximately 116 kb flanked by the two markers, pC43 at 8,655,291 bp and pC50 at 8,771,761 bp on 1DS (based on the *Ae. tauschii* reference genome Aet v4.0 (Luo *et al.* 2017)) (**Figure 4.1d**). Note that pC43 is an effective co-dominant marker to select *Lr42R*-carrying lines in both *Ae. tauschii* and bread wheat lines (**Table D.6**). The two markers are located on two genes that flank three other genes including an intact NLR gene (AET1Gv20040300), an NLR fragment (AET1Gv20040500), and a protein kinase (AET1Gv20040200).

Gene transfer confirmed that one NLR is *Lr42*

Our BSR-Seq result showed that all three candidate genes in the mapping interval were expressed in uninfected seedling leaves in both resistant and susceptible *Ae. tauschii* lines. Sequence comparison of the candidate genes between the resistant and susceptible lines using RNA-Seq data revealed that only the candidate AET1Gv20040300 showed polymorphisms in transcribed regions (**Figure D.2, Figure D.3**). We then amplified the full-length cDNA of AET1Gv20040300 from the resistant donor, TA2450, and the susceptible accession TA10132, and

confirmed polymorphisms between the two alleles (**Figure 4.2c**). Both alleles were separately transferred to a bread wheat cultivar 'Bobwhite'. In total, we obtained four independent positive transgenic T0 lines carrying the *Lr42* resistance allele (*Lr42R*) from TA2450 and two carrying the *lr42* susceptibility allele (*lr42S*) from TA10132. All T1 the transgenic lines were evaluated for leaf rust resistance.

Bobwhite carries the leaf rust resistance gene *Lr26* that confers resistance to many leaf rust *P. triticina* races (Germán and Kolmer 2012). We screened six *P. triticina* races and found that Bobwhite was susceptible to the race TFBJQ (**Table D.2**), which presumably overcame *Lr26* resistance, therefore TFBJQ was used as the inoculum to evaluate all the T1 transgenic lines for leaf rust resistance. Infection with the race TFBJQ found that all *Lr42R* transgenic lines gained high resistance and two *lr42S* transgenic lines were highly susceptible like Bobwhite (**Figure 4.2A**). Gene expression analysis showed that the *Lr42R* allele was expressed in all *Lr42R* transgenic lines and *lr42S* was expressed in both *lr42S* transgenic lines (**Figure 4.2B**). Note that expression of both *Lr42R* and *lr42S* were not detectable in Bobwhite. The phenotypic and gene expression data of transgenic lines confirmed that the NLR gene AET1Gv20040300 is *Lr42*.

Lr42 resistance allele rarely occurs in the *Ae. tauschii* collection

We amplified the *Lr42* homologs from 35 (failed to obtain genomic DNAs from 5 of the 40 accessions) out of 40 accessions from a minicore set of *Ae. tauschii* that captured >80% of genetic diversity of 549 *Ae. tauschii* accessions maintained at the Wheat Genetics Resource Center (WGRC) (Singh *et al.* 2019). *Ae. tauschii* has been classified to two major lineage groups: *Ae. tauschii* ssp. *tauschii* (Lineage 1, L1) and ssp. *strangulata* (Lineage 2, L2) (Singh *et al.* 2019) and 35 examined minicore accessions include 24 L1 and 11 L2 accessions. The *Lr42* donor TA2450

is an L2 accession. The *Lr42* gene failed to be amplified from 21/24 L1 accessions. Expected bands were amplified from 8/11 L2 accessions and 3 L1 accessions. Of them, 10 *Lr42* homologs were successfully sequenced (**Table D.1**). We also extracted intact *Lr42* homologs from the *Ae. tauschii* reference genome of the leaf rust susceptible accession TA10132. The homolog from TA10132 (*lr42s*) with the highest similarity and located in the mapping interval, which was used in the transgenic experiment is considered to be the allelic homolog of *Lr42*. Comparison of all these *Ae. tauschii* homologous sequences with *Lr42R* showed that 10 homologs from 10 *Ae. tauschii* minicore accessions are closer to the *Lr42R* as compared to non-allelic homologs in the reference genome, supporting that these 10 sequences from 10 accessions are allelic to *Lr42* (**Figure 4.3a**). Three seedling leaf rust resistant accessions TA1651, TA1667, TA2458 contained different *Lr42* haplotypes. Among these, TA1667 and TA2458 haplotypes were found in susceptible accessions, suggesting that the *Lr42* allelic homologs are not responsible for the leaf rust resistance in these two accessions. The phylogenetic analysis also indicated that the *Lr42* alleles are not completely separated in the two *Ae. tauschii* lineages (**Figure 4.3a**). Sequences of 11 *Lr42* allelic homologs, including *lr42S*, belonged to three major haplotypes, I, II, and III, represented by *lr42-TA2376*, *lr42-TA1605*, and *lr42-TA2536*, respectively (**Figure 4.3b**). Most sequences of the *Lr42R* allele can be found from these three haplotypes except for a segment in the LRR region (**Figure D.4, Table D.7**). Interestingly, the ~140 bp *Lr42R* unique segment in the LRR region (referred to as *Lr42R*-unique-segment hereafter) can be identified in a non-allelic *Lr42* homolog with 98% identity (the homolog 1D:7381846-7384626 has only 83.8% identity to *Lr42R*) from 1D subgenome of the wheat Chinese Spring (CS) reference genome (**Figure 4.3b, Figure D.5, Figure D.6**), implying that this unique sequence originated through either intragenic recombination or ectopic recombination. Beside the uniqueness of *Lr42R*, we also observed conserved sequences at

the beginning of the RX-CC domain, at the end of the NB-ARC domain, and at the beginning and the end of LRR (**Figure 4.3b**). Separate phylogenetic analysis using these domains (e.g., RX-CC, NB-ARC, and LRR) of the gene resulted in different phylogenetic relationships among these *Ae. tauschii* accessions further supported intragenic recombination occurred between *Lr42* haplotypes or ectopic recombination at some domains (**Figure D.4**).

In the *Ae. tauschii* reference genome, the susceptibility *lr42* allele (*lr42S*), four intact homologs, and four partial gene fragments were clustered within a 871 kb region (**Figure 4.3c**). Interestingly, homologous sequences with plus and minus orientations were physically separated into two regions, and homologs with the same orientation are more similar. The organization of the gene cluster indicated that *Lr42* homologs likely expanded independently in the two separate regions. The *Lr42* homologous clusters were also identified in 1A, 1B, 1D subgenomes of the hexaploid wheat variety, CS, 1A and 1B chromosomes of tetraploid emmer wheat, 1A and 1B chromosomes of durum wheat, as well as 1H of diploid barley (**Supplemental_file_4.1**). Only two homologs were identified in Brachypodium, a more distantly related species (**Figure 4.3c**). The results indicated that *Lr42* was derived from an ancient locus that has been maintained or expanded to result in a high copy number in barley and wheat species.

Lr42 exhibits effective resistance in wheat breeding programs

To identify which CIMMYT wheat lines contain the *Lr42* segment, both pedigree information and genotyping data via Genotyping-By-Sequencing (GBS) of 52,943 CIMMYT lines (Juliana *et al.* 2019) were used. We identified 14 *Lr42*-specific GBS tags (**Table D.5**). Of 5,121 genotyped CIMMYT lines with *Lr42*-carrying donors in their pedigree, 33.7% (1,724/5,121) were classified as *Lr42+* lines (**Figure 4.4a,** Supplemental_file_4.2). In contrast, only 2% (928/47,822)

of lines that were not expected to carry *Lr42* based on pedigree were categorized as *Lr42*+ lines. The two percent misclassified lines may reflect the false positive rate or could represent incorrect pedigrees or seed mixtures. In total, 2,924 out of 5,121 with an *Lr42* donor in the pedigree were categorized as without the *Lr42* segment (*Lr42*-) (**Figure 4.4a,** Supplemental_file_4.2).

Some *Lr42*+ and *Lr42*- wheat lines were phenotypically examined for leaf rust resistance and grain yield at CIMMYT (Juliana *et al.* 2019). Comparison between *Lr42*+ and *Lr42*- wheat lines from the breeding population supported that the *Lr42* segment is highly associated with seedling resistance to the leaf rust race MBJ/SP (Lan *et al.* 2014), and moderately associated with resistance at the adult stage to leaf rust (**Figure 4.4b**). Without leaf rust infection, grain yield traits of *Lr42*+ and *Lr42*- lines were not significantly different, indicative of no significant yield boost or penalty directly imposed by the *Lr42* resistance segment from *Ae. tauschii* (**Table D.8**).

Diagnostic markers for *Lr42* genotyping

To facilitate marker-assisted selection of *Lr42* in breeding, we developed an effective co-dominant KASP marker pC43 that is located 46kb from the *Lr42* gene for selection of *Lr42*-carrying lines in both *Ae. tauschii* and bread wheat lines. We have also designed and validated two markers, Lr42-pD1 and Lr42-pD2, on the *Lr42* gene to distinguish the presence or absence of the *Lr42* resistance allele in wheat (**Table D.6, Table D.9**). These markers will facilitate deployment of *Lr42* in wheat breeding programs through precise marker-assisted selection.

Discussion

We employed an efficient mapping strategy using diploid *Ae. tauschii* populations to clone the broadly effective leaf rust resistance gene *Lr42* by taking advantage of the newly constructed *Ae. tauschii* reference genome (Luo *et al.* 2017), high-throughput sequencing technology, and the

optimized genetic analysis strategy, BSR-Seq (Liu *et al.* 2012). Using susceptible and resistant bulks, BSR-Seq enabled simultaneous high-density SNP discovery and genotyping to map the genomic region that contains *Lr42*. Further fine-mapping delimited the gene interval to approximately 116 kb and revealed an expressed candidate NLR gene, AET1Gv20040300, for *Lr42*. The causal gene was confirmed by gain-of-resistance *via* gene transfer to a susceptible hexaploid wheat cultivar.

The cloning of *Lr42* added a new member to at least 12 cloned wheat rust resistance NLR genes (**Figure D.7**) (Huang *et al.* 2003; Feuillet *et al.* 2003; Cloutier *et al.* 2007; Periyannan *et al.* 2013; Saintenac *et al.* 2013; Liu *et al.* 2014; Mago *et al.* 2015; Steuernagel *et al.* 2016; Thind *et al.* 2017; Zhang *et al.* 2017, 2019). NLR functions as an intracellular sensor of pathogen signals and/or as an executor to induce localized cell death, the hypersensitive immune response. NLRs exerting both functions were recently referred to as singleton NLRs, such as *Mla* (Saur *et al.* 2019) and *Sr50* (Chen *et al.* 2017). Some other NLRs function in a pair: sensor NLR recognizing the pathogen and helper (or executor) NLR initiating immune signaling. The paradigm of NLR networks consisting of a number of sensor NLRs and helper NLRs to modulate immune responses was also proposed (Adachi *et al.* 2019). An NLR gene in monocots generally consists of an N-terminal coiled-coil (CC) domain, the central NB-ARC domain, and a C-terminal leucine-rich LRR domain. Recent protein structure studies of an Arabidopsis NLR gene product, ZAR1, revealed that a pentameric wheel-like NLR resistosome is assembled upon activation by the pathogen. The funnel-shaped structure formed from the N terminal α helices at the CC domain is hypothesized to directly compromise plasma membrane integrity and induce cell death (Wang *et al.* 2019a; b). Interestingly, a MADA motif (MADAxVSFxVxKLxxLLxxEx, where x represents non-conserved amino acids) conserved among helper NLRs and singleton NLRs but not sensor

NLRs, was identified on the CC domain. *Lr42* has a typical NLR structure and contains a homologous domain "MAEAVVGQLVVTLGEALAKEA", which is the most similar to the MADA motif among all known wheat rust resistance NLRs (**Table D.10**). This implies that *Lr42* is more likely to be a singleton NLR or a helper NLR, not a sensor NLR.

Lr42R is apparently a new allele at an ancient locus. Homologs of *Lr42* were detected in the distant wheat relative Brachypodium (**Figure 4.3c**), which diverged from the Triticeae (wheat, rye, barley) lineage 32-39 MYA (International Brachypodium Initiative 2010). Nevertheless, 34/35 samples in the *Ae. tauschii* minicore have been excluded to carry a resistance *Lr42* allele, suggesting that *Lr42* is rare and, likely, of recent origin. The variation in LRR repeat numbers among *Lr42* alleles indicated that unequal crossovers could have occurred within the LRR domain (Richter and Ronald 2000). In addition, intragenic recombination as was documented for *Lr21* (Huang *et al.* 2009), or even ectopic recombination, may also have played a role in the origin of *Lr42* allele. Indeed, the unique LRR sequence of the *Lr42* allele can be identified in a non-allelic region in the subgenome 1D of CS, supporting the potential role of ectopic recombination in the origin of the *Lr42* resistance allele.

The phenotypic expression of resistance in *Lr42* lines depends on several factors. Although no leaf rust isolates have shown full virulence to *Lr42*, some isolates showed lower infection types than others on KS93U50, an *Lr42* resistant selection from KS91WGRC11 (Sun *et al.* 2010). The resistance reaction of the diploid *Ae. tauschii* TA2450 donor accession is consistently very strong, ranging from a hypersensitive fleck (Infection Type (IT) = ;) to flecks with tiny pustules surrounded by necrosis (IT = ;1-) (**Figure 4.1a**). However, the reaction of nontransgenic hexaploid *Lr42*--containing lines ranged from flecks and small pustules surrounded by necrosis (IT = ;1) to medium-sized pustules surrounded by chlorosis (IT = 2+) (Cox *et al.* 1994a; Sun *et al.* 2010). The

reduced expression of introgressed resistance in hexaploid bread wheat compared to diploid donors is a frequently observed phenomenon (Wulff and Moscou 2014). In contrast, the reaction of the transgenic hexaploid derivatives was very strong, ranging from a hypersensitive fleck (IT = ;) to flecks with tiny pustules (IT = ;1-) (**Figure 4.2a**). The improved performance of the transgenic versus nontransgenic hexaploid lines may be due to the strong maize ubiquitin promoter that was used in the transgenics. The very strong resistance of the transgenic hexaploid *Lr42* lines bodes well for its utility in transgenic cassettes. Plant age and/or environment may also influence *Lr42* resistance. Adult plants in the field showed much stronger *Lr42* resistance than greenhouse-grown seedlings (Martin *et al.* 2003).

The undefeated status of *Lr42* raised the possibility that it might be a more durable type of resistance gene. However, elucidation of the NLR structure of *Lr42* indicates that the mechanism of resistance is typical effector-triggered immunity (ETI). ETI is usually not durable because the rust pathogen can become virulent by loss of the corresponding avirulence factor (effector) that triggers the hypersensitive resistance response. *Lr42* is currently deployed mainly in wheats from CIMMYT that contain combinations of durable adult plant resistance (APR) genes to leaf rust (**Table D.11**). This may have reduced the selection pressure on the pathogen population to overcome *Lr42*. The CIMMYT wheat breeding pipeline has many *Lr42*-containing breeding lines in a background with high levels of APR to leaf rust (**Table D.9**). Effective gene stewardship will require breeders to release *Lr42* only in varieties with strong combinations of other leaf rust resistance genes.

Previous field trials showed that the *Lr42* introgression contributed to large increases in yield and kernel weight in Oklahoma (Martin *et al.* 2003). We used GBS markers to classify 5,121 CIMMYT breeding lines that had *Lr42* in the pedigree. Some of advanced lines positive for *Lr42*

were compared to their counterparts without *Lr42*. We were able to detect a very large effect of *Lr42* on leaf rust ratings at the seedling stage, but only a moderate effect on severity at the adult stage in the field probably because most CIMMYT lines also had a high level of APR that kept disease severities low (**Figure 4.4c**). In a QTL analysis of a highly resistant CIMMYT line Quaiu 3, Basnet et al (2014) were able to separate the effect of *Lr42* from other resistance genes. *Lr42* explained 32% of the phenotypic variation and limited disease severity in the field to a maximum of 40%. *Lr42* combined very well with *Lr46* and *QLr.tam-3D* to achieve near immunity to leaf rust in Quaiu 3 (Basnet *et al.* 2014). We did not detect a direct or indirect impact of *Lr42* on yield and other grain quality traits in the *Lr42*+ lines, also probably due to a high level of APR in most CIMMYT lines.

KS91WGRC11 may be common in CIMMYT pedigrees because it contributes resistance to stem rust and stripe rust in addition to leaf rust. KS91WGRC11 carries the *SrTmp* stem rust resistance gene on chromosome 6DS from the Century parent (Lopez-Vera *et al.* 2014). We also documented a hidden introgression in the WGRC germplasm. Recently, a stripe rust resistance NLR gene *YrAS2388* originating from chromosome 4D of an *Ae. tauschii* accession was cloned and TA2450 was found to carry the resistance allele (Zhang *et al.* 2019). We amplified the *YrAS2388* gene from TA2450 and confirmed that the sequence is identical to the reference resistance allele reported. We also found that KS91WGRC11 carries the *YrAS2388* resistance allele from TA2450, which implies that the *YrAS2388* resistance allele has been introduced to germplasm in CIMMYT and many other breeding programs. Given the limited backcrosses to Century, KS91WGRC11 is expected to harbor additional genomic segments from *Ae. tauschii* that might contribute valuable genetic diversity to future cultivars. Our results point to the need for *in situ* conservation of robust populations of native wild species for enhancing crop biodiversity so

that rare alleles such as *Lr42* reported here can evolve and be conserved for future crop improvement.

Conclusion

The *Lr42* gene from the wheat wild relative *Aegilops tauschii* confers resistance to all leaf rust races tested to date. Through bulked segregant RNA-Seq (BSR-Seq) mapping and further fine mapping, we identified an *Lr42* candidate gene, which encodes nucleotide-binding site leucine-rich repeat (NLR). Transformation of the candidate gene to a leaf rust susceptible wheat cultivar markedly enhanced the disease resistance, confirming the candidate NLR gene is the *Lr42* gene. Cloning of *Lr42* expands the repertoire of cloned rust resistance genes, as well as provides precise diagnostic DNA markers for wheat improvement.

Data availability

BSR-Seq raw sequencing data generated during the current study are available in the NCBI BioProject under accession number of PRJNA604114.

References

Adachi H., L. Derevnina, and S. Kamoun, 2019 NLR singletons, pairs, and networks: evolution, assembly, and regulation of the intracellular immunoreceptor circuitry of plants. Curr. Opin. Plant Biol. 50: 121–131.

Altschul S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Avni R., M. Nave, O. Barad, K. Baruch, S. O. Twardziok, *et al.*, 2017 Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357: 93–97.

Basnet B. R., R. P. Singh, S. A. Herrera-Foessel, A. M. H. Ibrahim, J. Huerta-Espino, *et al.*, 2013 Genetic analysis of adult plant resistance to yellow rust and leaf rust in common spring wheat Quaiu 3. Plant Dis. 97: 728–736.

Basnet B. R., R. P. Singh, A. M. H. Ibrahim, S. A. Herrera-Foessel, J. Huerta-Espino, *et al.*, 2014 Characterization of Yr54 and other genes associated with adult plant resistance to yellow rust and leaf rust in common wheat Quaiu 3. Mol. Breed. 33: 385–399.

Bej A., B. R. Sahoo, B. Swain, M. Basu, P. Jayasankar, *et al.*, 2014 LRRsearch: An asynchronous server-based application for the prediction of leucine-rich repeat motifs and an integrative database of NOD-like receptors. Comput. Biol. Med. 53: 164–170.

Bernardo A., P. St. Amand, H. Q. Le, Z. Su, and G. Bai, 2020 Multiplex restriction amplicon sequencing: a novel next-generation sequencing-based marker platform for high-throughput genotyping. Plant Biotechnol. J. 18: 254–265.

Chen J., N. M. Upadhyaya, D. Ortiz, J. Sperschneider, F. Li, *et al.*, 2017 Loss of AvrSr50 by somatic exchange in stem rust leads to virulence for Sr50 resistance in wheat. Science 358: 1607–1610.

Christensen A. H., and P. H. Quail, 1996 Ubiquitin promoter-based vectors for high-level expression of selectable and/or screenable marker genes in monocotyledonous plants. Transgenic Res. 5: 213–218.

Cloutier S., B. D. McCallum, C. Loutre, T. W. Banks, T. Wicker, *et al.*, 2007 Leaf rust resistance gene Lr1, isolated from bread wheat (Triticum aestivum L.) is a member of the large psr567 gene family. Plant Mol. Biol. 65: 93–106.

Cox T. S., W. J. Raupp, and B. S. Gill, 1994a Leaf rust-resistance genes Lr41, Lr42, and Lr43 transferred from Triticum tauschii to common wheat. Crop Sci. 34: 339–343.

Cox T. S., R. G. Sears, B. S. Gill, and E. N. Jellen, 1994b Registration of KS91WGRC11, KS92WGRC15, KS92WGRC23 leaf rust-resistant hard red winter wheat germplasms. Crop Sci. 34: 546–547.

Feuillet C., S. Travella, N. Stein, L. Albar, A. Nublat, *et al.*, 2003 Map-based isolation of the leaf rust disease resistance gene Lr10 from the hexaploid wheat (Triticum aestivum L.) genome. Proc. Natl. Acad. Sci. U. S. A. 100: 15253–15258.

Germán S. E., and J. A. Kolmer, 2012 Leaf rust resistance in selected Uruguayan common wheat cultivars with early maturity. Crop Sci. 52: 601–608.

Gill H. S., C. Li, J. S. Sidhu, W. Liu, D. Wilson, *et al.*, 2019 Fine Mapping of the Wheat Leaf Rust Resistance Gene Lr42. Int. J. Mol. Sci. 20. https://doi.org/10.3390/ijms20102445

Huang L., S. A. Brooks, W. Li, J. P. Fellers, H. N. Trick, *et al.*, 2003 Map-based cloning of leaf rust resistance gene Lr21 from the large and polyploid genome of bread wheat. Genetics 164: 655–664.

Huang L., S. Brooks, W. Li, J. Fellers, J. C. Nelson, *et al.*, 2009 Evolution of New Disease Specificity at a Simple Resistance Locus in a Crop–Weed Complex: Reconstitution of the Lr21 Gene in Wheat. Genetics 182: 595–602.

Huerta-Espino J., R. P. Singh, S. Germán, B. D. McCallum, R. F. Park, *et al.*, 2011 Global status of wheat leaf rust caused by Puccinia triticina. Euphytica 179: 143–160.

International Brachypodium Initiative, 2010 Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463: 763–768.

International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators:, R. Appels, K. Eversole, C. Feuillet, *et al.*, 2018 Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361. https://doi.org/10.1126/science.aar7191

Juliana P., J. Poland, J. Huerta-Espino, S. Shrestha, J. Crossa, *et al.*, 2019 Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat. Genet. 51: 1530–1539.

Kolmer J. A., 1996 Genetics of resistance to wheat leaf rust. Annu. Rev. Phytopathol. 34: 435–455.

Kolmer J. A., 2019 Virulence of Puccinia triticina, the Wheat Leaf Rust Fungus, in the United States in 2017. Plant Dis. 103: 2113–2120.

Krattinger S. G., E. S. Lagudah, W. Spielmeyer, R. P. Singh, J. Huerta-Espino, *et al.*, 2009 A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. Science 323: 1360–1363.

Krzywinski M., J. Schein, I. Birol, J. Connors, R. Gascoyne, *et al.*, 2009 Circos: an information aesthetic for comparative genomics. Genome Res. 19: 1639–1645.

Lan C. X., R. P. Singh, J. Huerta-Espino, V. Calvo-Salazar, and S. A. Herrera-Foessel, 2014 Genetic analysis of resistance to leaf rust and stripe rust in wheat cultivar Francolin# 1. Plant Dis. 98: 1227–1234.

Letunic I., and P. Bork, 2019 Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47: W256–W259.

Li W., and A. Godzik, 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.

Liu S., C.-T. Yeh, H. M. Tang, D. Nettleton, and P. S. Schnable, 2012 Gene mapping via bulked segregant RNA-Seq (BSR-Seq). PLoS One 7: e36406.

Liu Z., R. L. Bowden, and G. Bai, 2013 Molecular Markers for Leaf Rust Resistance Gene Lr42 in Wheat. Crop Sci. 53: 1566–1570.

Liu W., M. Frick, R. Huel, C. L. Nykiforuk, X. Wang, *et al.*, 2014 The stripe rust resistance gene Yr10 encodes an evolutionary-conserved and unique CC-NBS-LRR sequence in wheat. Mol. Plant 7: 1740–1755.

Lopez-Vera E. E., S. Nelson, R. P. Singh, B. R. Basnet, S. D. Haley, *et al.*, 2014 Resistance to stem rust Ug99 in six bread wheat cultivars maps to chromosome 6DS. Theor. Appl. Genet. 127: 231–239.

Luo M.-C., Y. Q. Gu, D. Puiu, H. Wang, S. O. Twardziok, *et al.*, 2017 Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature 551: 498–502.

Maccaferri M., N. S. Harris, S. O. Twardziok, R. K. Pasam, H. Gundlach, *et al.*, 2019 Durum wheat genome highlights past domestication signatures and future improvement targets. Nat. Genet. 51: 885–895.

Mago R., P. Zhang, S. Vautrin, H. Šimková, U. Bansal, *et al.*, 2015 The wheat Sr50 gene reveals rich diversity at a cereal disease resistance locus. Nat Plants 1: 15186.

Marchler-Bauer A., Y. Bo, L. Han, J. He, C. J. Lanczycki, *et al.*, 2017 CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45: D200–D203.

Martin J. N., B. F. Carver, R. M. Hunger, and T. S. Cox, 2003 Contributions of leaf rust resistance and awns to agronomic and grain quality performance in winter wheat. Crop Sci. 43: 1712–1717.

Mascher M., H. Gundlach, A. Himmelbach, S. Beier, S. O. Twardziok, *et al.*, 2017 A chromosome conformation capture ordered sequence of the barley genome. Nature 544: 427–433.

McKenna A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Moore J. W., S. Herrera-Foessel, C. Lan, W. Schnippenkoetter, M. Ayliffe, *et al.*, 2015 A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. Nat. Genet. 47: 1494–1498.

Periyannan S., J. Moore, M. Ayliffe, U. Bansal, X. Wang, *et al.*, 2013 The gene Sr33, an ortholog of barley Mla genes, encodes resistance to wheat stem rust race Ug99. Science 341: 786–788.

Pfeifer B., U. Wittelsbürger, S. E. Ramos-Onsins, and M. J. Lercher, 2014 PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31: 1929–1936.

Richter T. E., and P. C. Ronald, 2000 The evolution of disease resistance genes. Plant Mol. Biol. 42: 195–204.

Saintenac C., W. Zhang, A. Salcedo, M. N. Rouse, H. N. Trick, *et al.*, 2013 Identification of wheat gene Sr35 that confers resistance to Ug99 stem rust race group. Science 341: 783–786.

Saur I. M., S. Bauer, B. Kracher, X. Lu, L. Franzeskakis, *et al.*, 2019 Multiple pairs of allelic MLA immune receptor-powdery mildew AVRA effectors argue for a direct recognition mechanism. Elife 8: e44471.

Singh N., S. Wu, V. Tiwari, S. Sehgal, J. Raupp, *et al.*, 2019 Genomic analysis confirms population structure and identifies inter-lineage hybrids in Aegilops tauschii. Front. Plant Sci. 10: 9.

Steuernagel B., S. K. Periyannan, I. Hernández-Pinzón, K. Witek, M. N. Rouse, *et al.*, 2016 Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. Nat. Biotechnol. 34: 652–655.

Sun X., G. Bai, B. F. Carver, and R. Bowden, 2010 Molecular mapping of wheat leaf rust resistance gene Lr42. Crop Sci. 50: 59–66.

Tassy C., and P. Barret, 2017 Biolistic Transformation of Wheat. Methods Mol. Biol. 1679: 141–152.

Thind A. K., T. Wicker, H. Šimková, D. Fossati, O. Moullet, *et al.*, 2017 Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. Nat. Biotechnol. 35: 793–796.

Tian B., M. Navia-Urrutia, Y. Chen, J. Brungardt, and H. N. Trick, 2019 Biolistic Transformation of Wheat. Methods Mol. Biol. 1864: 117–130.

Wang J., M. Hu, J. Wang, J. Qi, Z. Han, *et al.*, 2019a Reconstitution and structure of a plant NLR resistosome conferring immunity. Science 364: eaav5870.

Wang J., J. Wang, M. Hu, S. Wu, J. Qi, *et al.*, 2019b Ligand-triggered allosteric ADP release primes a plant NLR complex. Science 364: eaav5868.

Wu T. D., and S. Nacu, 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873–881.

Wulff B. B. H., and M. J. Moscou, 2014 Strategies for transferring resistance into wheat: from wide crosses to GM cassettes. Front. Plant Sci. 5: 692.

Zhang D., R. L. Bowden, J. Yu, B. F. Carver, and G. Bai, 2014 Association analysis of stem rust resistance in U.S. winter wheat. PLoS One 9: e103747.

Zhang W., S. Chen, Z. Abate, J. Nirmala, M. N. Rouse, *et al.*, 2017 Identification and characterization of , a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. Proc. Natl. Acad. Sci. U. S. A. 114: E9483–E9492.

Zhang C., L. Huang, H. Zhang, Q. Hao, B. Lyu, *et al.*, 2019 An ancestral NB-LRR with duplicated 3'UTRs confers stripe rust resistance in wheat and barley. Nat. Commun. 10: 4023.

Figures

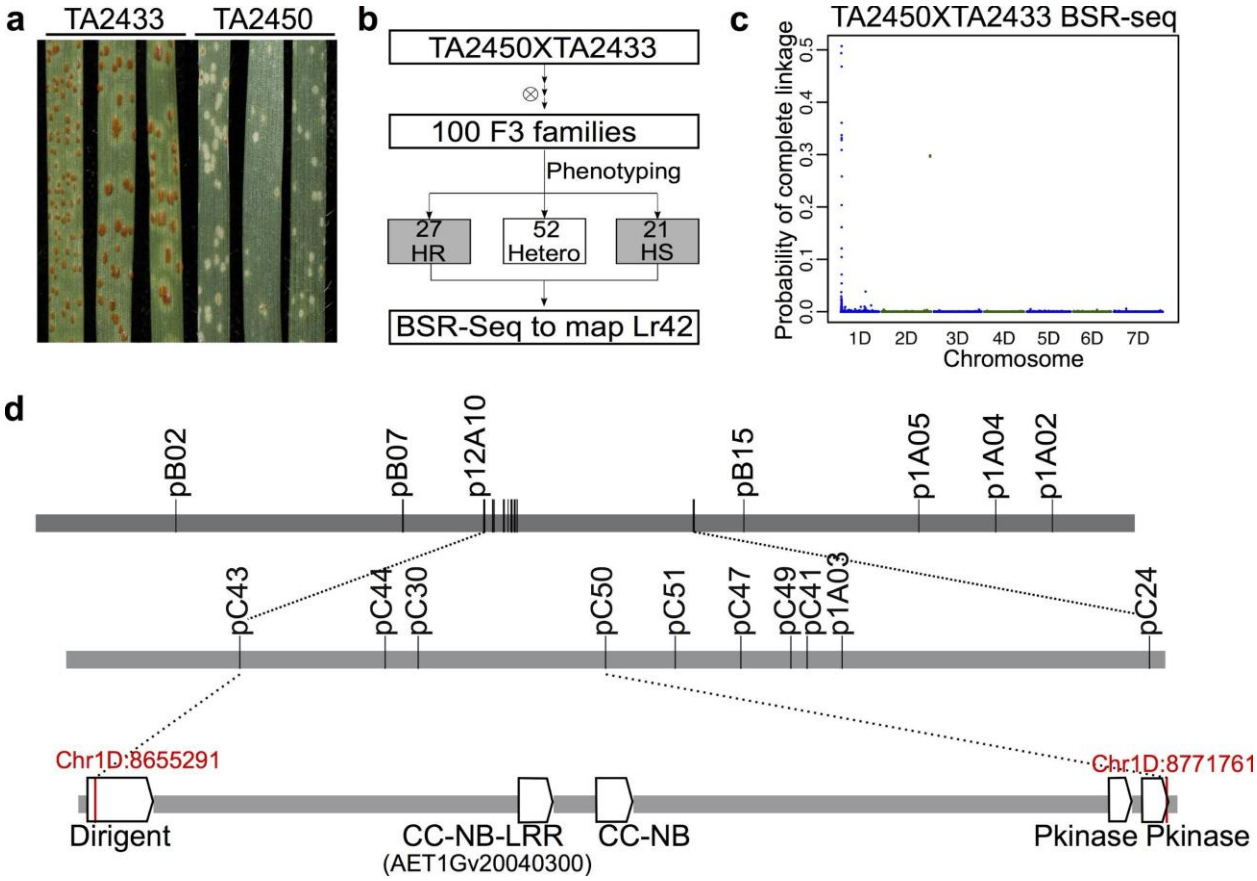

Figure 4.1 Genetic mapping of the *Lr42* gene

(**a**) Phenotype of *Ae. tauschii* accessions TA2433 (*Lr42*-) was susceptible (Infection Type = 33+) and TA2450 (*Lr42*+) was hypersensitive fleck (Infection Type = ; to ;1-) at the seedling stage upon inoculation with race PNMRJ. (**b**) Genetic mapping *Lr42* genes via BSR-seq. In total, we phenotyped 100 $F_{2:3}$ families (15 individuals for each family) from the cross TA2450 x TA2433 and identified 27 homozygous resistance (HR) families, 52 heterozygous (Hetero) families, and 21 homozygous susceptible (HS) families. Bulked segregant analysis via RNA-Seq (BSR-Seq) of 21 HR and 21 HS seedling pools was employed to map the *Lr42* gene. (**c**) For each variant identified from RNA-Seq, the probability of complete linkage between the variant and the *Lr42* gene was plotted versus the chromosomal position of the variant. (**d**) The first track shows some single-nucleotide polymorphisms (SNPs) identified from BSR-Seq. The second track shows KASP markers that were developed for *Lr42* fine-mapping. The two flanking markers pC43 and pC50 delineate the fine-mapping interval. The physical positions on the *Ae. tauschii* reference genome (Aet v4.0) of the KASP marker SNPs are shown in the third track. Annotated genes in the interval are represented by open boxes.

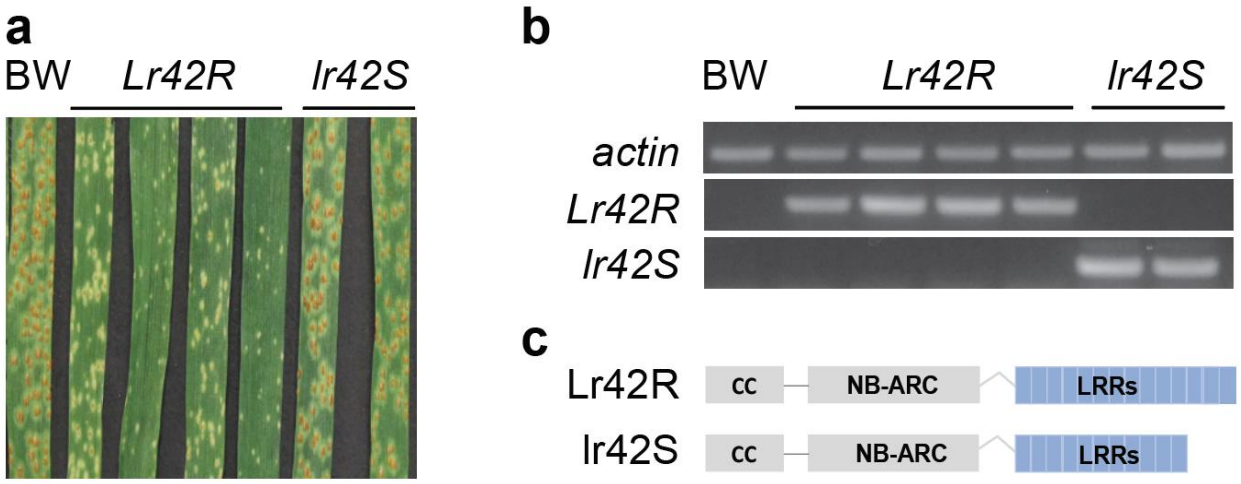

Figure 4.2 Validation of the *Lr42* candidate gene via transformation

(**a**) Phenotype of Bobwhite (BW, transgenic background) was susceptible (Infection Type = 3), four independent events of *Lr42* resistance allele (*Lr42R*) was hypersensitive fleck (Infection Type = ; to ;1-), and three independent events of *Lr42* susceptibility allele (*lr42S*) was susceptible (Infection Type = 3) upon inoculation with race TFBJG. Transgenic plants are in T1 generation. *Lr42R* is from accession TA2450, and *lr42S* is from accession TA10132. (**b**) RT-PCR of the *Lr42* gene of Bobwhite (BW) and transgenic plants from the T1 generation. Resistant and susceptible alleles were amplified with primer sets Lr42-qRT-F5/R5 and lr42_1F/R (**Table D.4**). (C) The protein structures of the *Lr42R* and *lr42S* alleles. Note that *lr42S* has fewer LRR repeats compared with *Lr42R*.

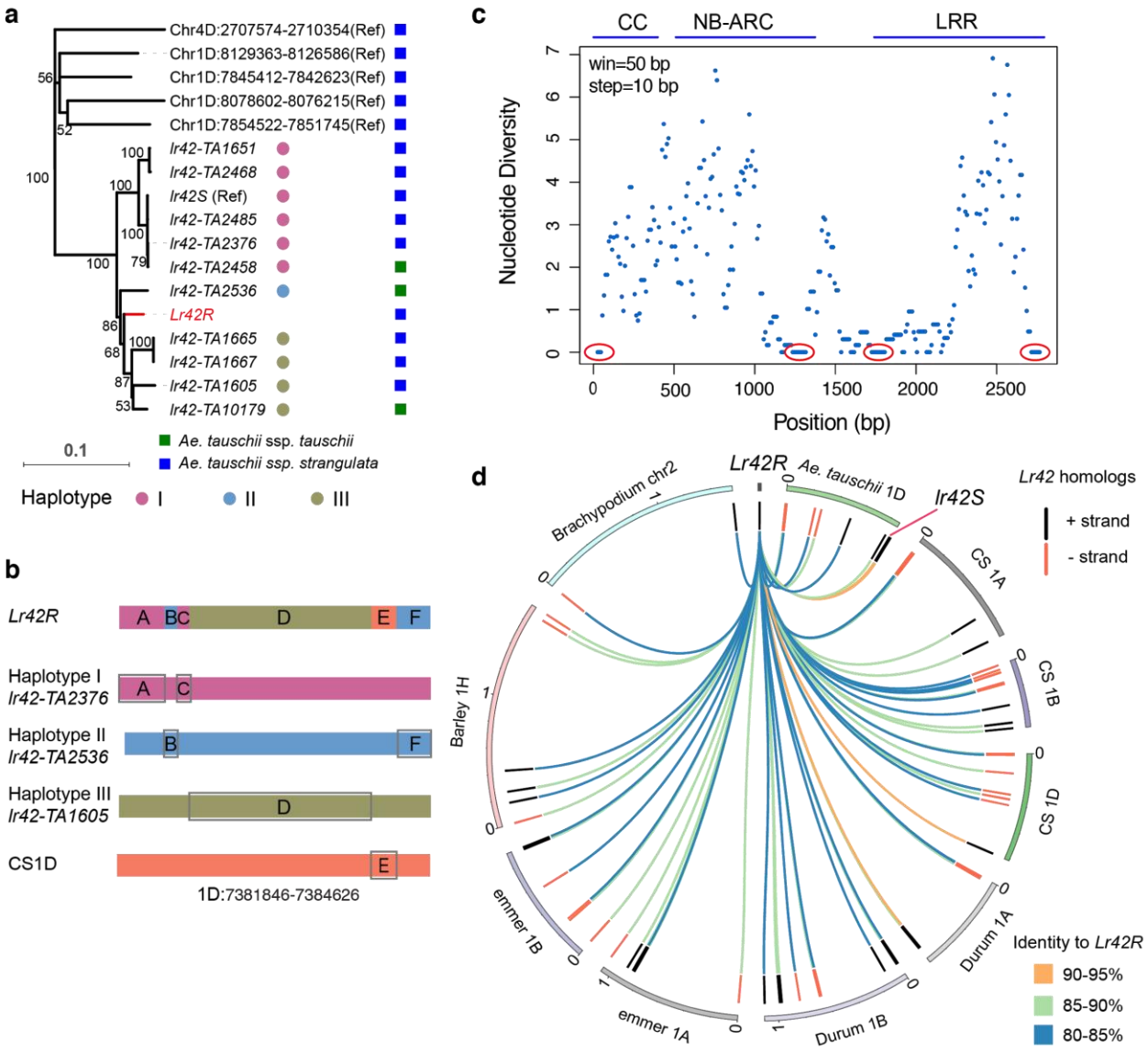

Figure 4.3 Homologs of *Lr42* in *Ae. tauschii* and closely related species

(**a**) Phylogenetic tree of intact *Lr42* homologs from *Ae. tauschii*. The *Lr42R* is the resistance allele. The *lr42S* allele is a susceptibility allele from the reference accession TA10132. The other 10 alleles (signified by *lr42*-accession) were from the *Ae. tauschii* minicore set. Solid colored circles represent *Lr42* haplotypes. Squares indicate lineages of accessions. Bootstraps are labeled on the tree. (**b**) Haplotype analysis. Eleven *Lr42* alleles that do not include *Lr42R* are clustered and grouped into three major haplotypes, represented by three alleles from TA2376, TA2536, and TA1605. Each sequence block (A to E) of *Lr42R* indicates the best hit with at least 95% identity to the block with the same letter on three haplotypes and a sequence fragment from 1D subgenome of CS. (**c**) The nucleotide diversity of the twelve *Lr42* alleles in 50-bp windows scanned on the gene with the step size of 10 bp. Each dot represents a nucleotide diversity of a 50-bp window versus the middle position of the window. Conserved regions with very low diversity are

highlighted by red ovals. **(d)** Circos view of *Lr42* homologs. *Lr42* clusters include *Lr42* homologs (at least 1 kb match and 79% identity) on *Ae. tauschii* 1D (7,842,623-8,713,757 bp), bread wheat CS 1A (8,612,188-9,546,340 bp), 1B (9,547,814-10,062,691 bp), 1D (7,056,154-7,871,394 bp), durum wheat 1A (8,461,802-9,214,843), 1B (8,108,317-9,210,082), Wild emmer 1A (10,412,667-11,483,838 bp), 1B (11,947,692-12,868,059 bp), Barley 1H (3,492,357-5216559 bp), and Brachypodium chromosome 2 (38,092,538-39,649,528 bp). The beginning of each cluster was adjusted to 0. The 1s on the cluster track represent 1 Mb positions. The identity between each homolog and *Lr42* is color-coded. The red line points at the position of the *Lr42* susceptible allele on the *Ae. tauschii* reference genome.

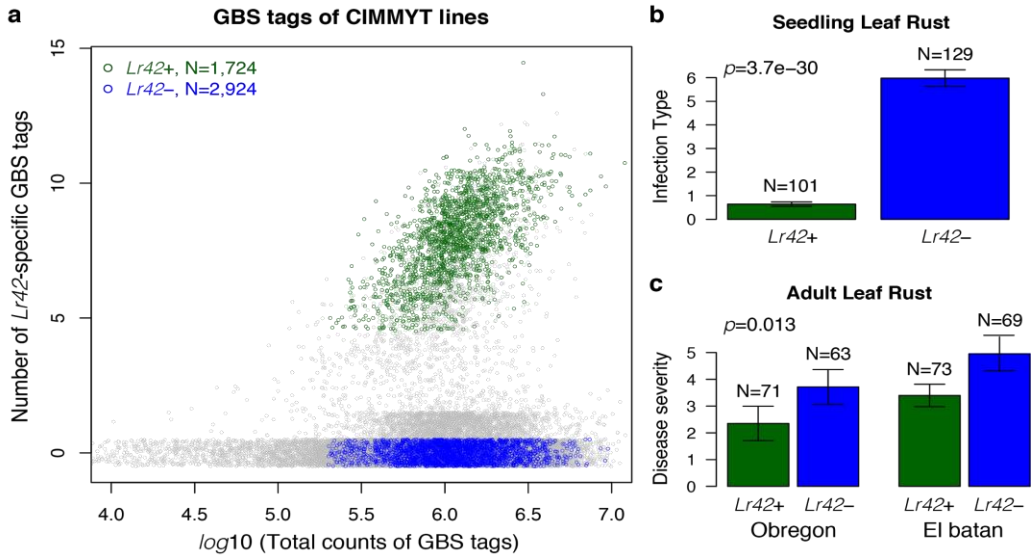

Figure 4.4 Introgression of the *Lr42* segment in CIMMYT wheat lines

(**a**) Each point represents a wheat line with the log10 of total count of GBS tags per line on the x-axis and the number of *Lr42*-specific GBS tags on the y-axis. Green and blue colors signify *Lr42+* and *Lr42-* *Lr42*-introgressed lines, respectively. Other lines that are either not *Lr42*-introgressed lines or not confidently categorized into *Lr42+* or *Lr42-* are gray colored. Numbers of *Lr42*-specific GBS tags were jittered with the factor 0.25 for better visualization of data density. (**b, c**) Using disease phenotypic data collected at CIMMYT, statistical comparisons of disease infection types between *Lr42+* and *Lr42-* lines were performed for seedling stage leaf rust with t-test and for adult stage resistance with ANOVA. Two field locations in Mexico for adult stage leaf rust phenotyping are labelled. The Stakman 0-4 scale was linearized to a 0–9 scale for seedlings (Methods). The percent disease severity on adult plant flag leaves was rated on the Cobb scale. Bars represent standard deviation of means of infection types or disease severity.

List of supplemental files

Supplemental_file_4.1 Homologs information in *Lr42* clusters

Supplemental_file_4.2 Presence of *Lr42* segment in CIMMYT lines

Appendix A - Copyright Information

This section contains the copyright information in the dissertation.

Chapter 2 in this dissertation is reprinted from Lin *et al.*, 2020 Chromosome-level Genome Assembly of a Regenerable Maize Inbred Line A188. bioRxiv 2020.09.09.289611. The authors hold the copyright of the content.

The research work in the other chapters has not been published yet. As the copyright holder, we reserve the right to publish it first. Permission is required from the authors before using this unpublished content from this dissertation.

Appendix B - Supplemental data of Chapter 2

Supplemental Figures

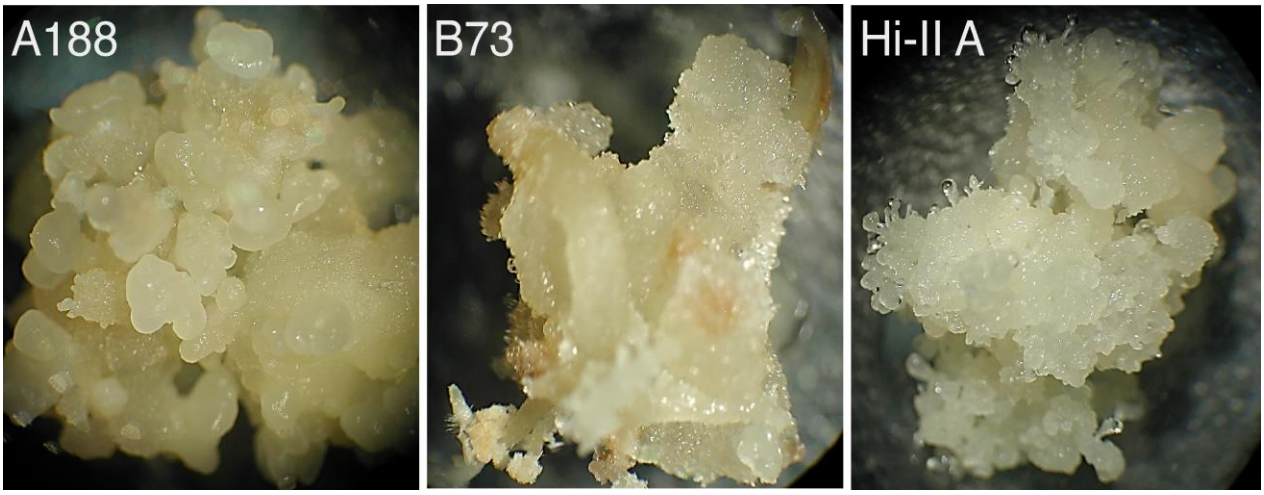

Figure B.1 Calli from immature embryos of A188, B73, and Hi-II A. a, **b**, **c**) White, compact, and nodulated somatic embryos and embryogenic callus from A188 **(a)**, brownish, loose and non-embryogenic callus from B73 **(b)**, and white, friable and embryogenic callus from Hi-II A (**c**) after 28 days of culture in callus induction media.

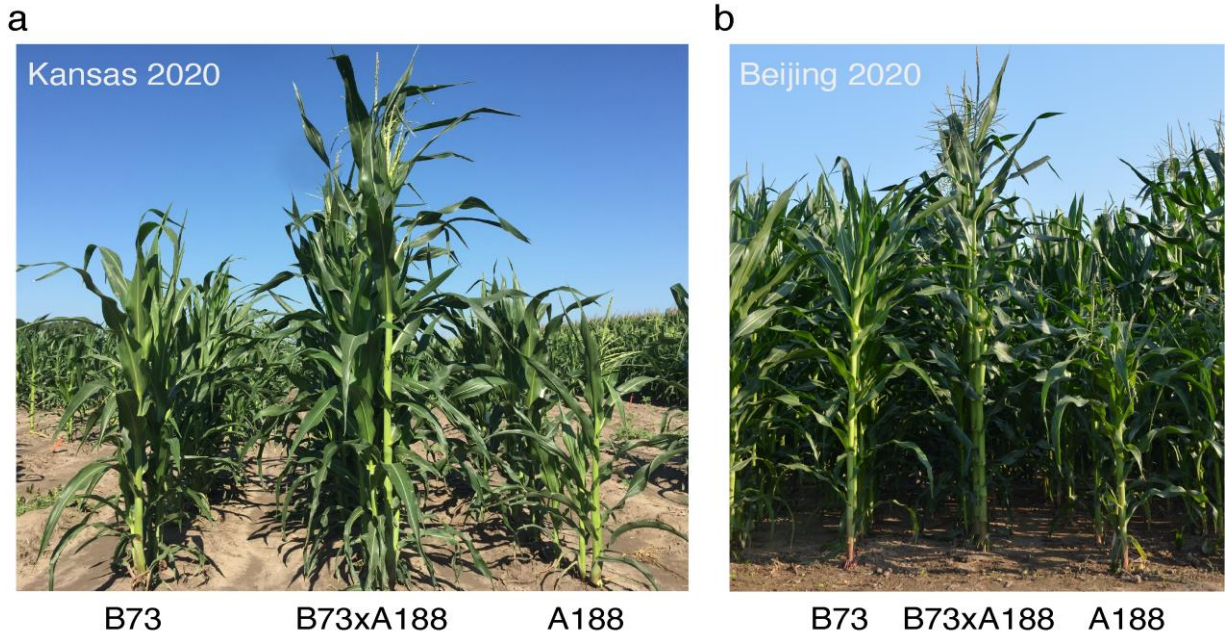

Figure B.2 Phenotypes of B73, F1, and A188. a,b). Plants from 2020 summer nursery in Kansas (**a**) and Beijing (**b**).

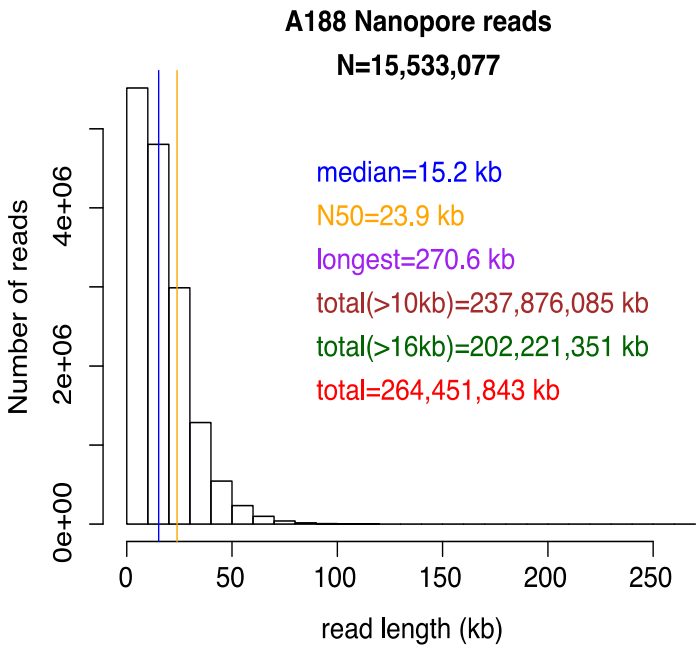

Figure B.3 Histogram of lengths of Nanopore raw reads. MinION flowcells (N=31) were used to produce Nanopore reads for the A188 genome assembly, producing >264 Gb total sequences. The median and N50 of read lengths are indicated by blue and red vertical lines, respectively.

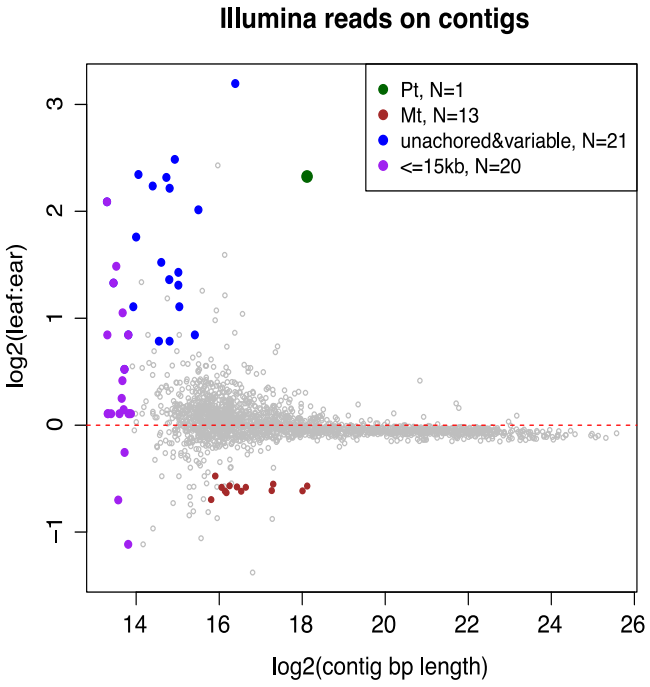

Figure B.4 Contig filtering based on read depths. Log2 values of Illumina read depth ratios of seedling leaf to ear samples, log2(leaf:ear), were determined for each contig. Contigs that were not anchored to B73Ref4 and showed a high variability from 0 of log2(leaf:ear) and contigs less than 15 kb were discarded. The chloroplast contig (pt) and mitochondrial contigs (mt) were replaced by the A188 chloroplast complete sequence (Genbank accession KF241980.1) and the A188 mitochondrion complete sequence (Genbank accession DQ490952.1), respectively.

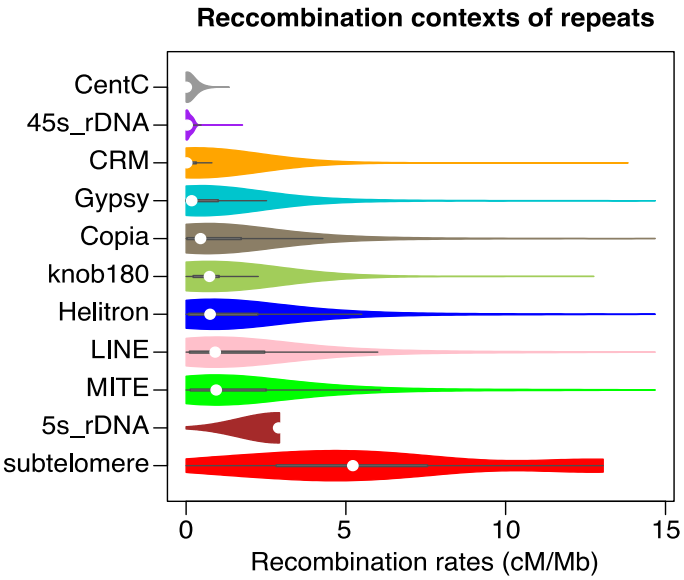

Figure B.5 Recombination of contexts around repeats. For each repeat type, the recombination rate (cM/Mb) of the surrounding 1 Mb context was estimated for each sequence on chromosomes. A violin plot of all sequences of each type was plotted with a dot to represent the median.

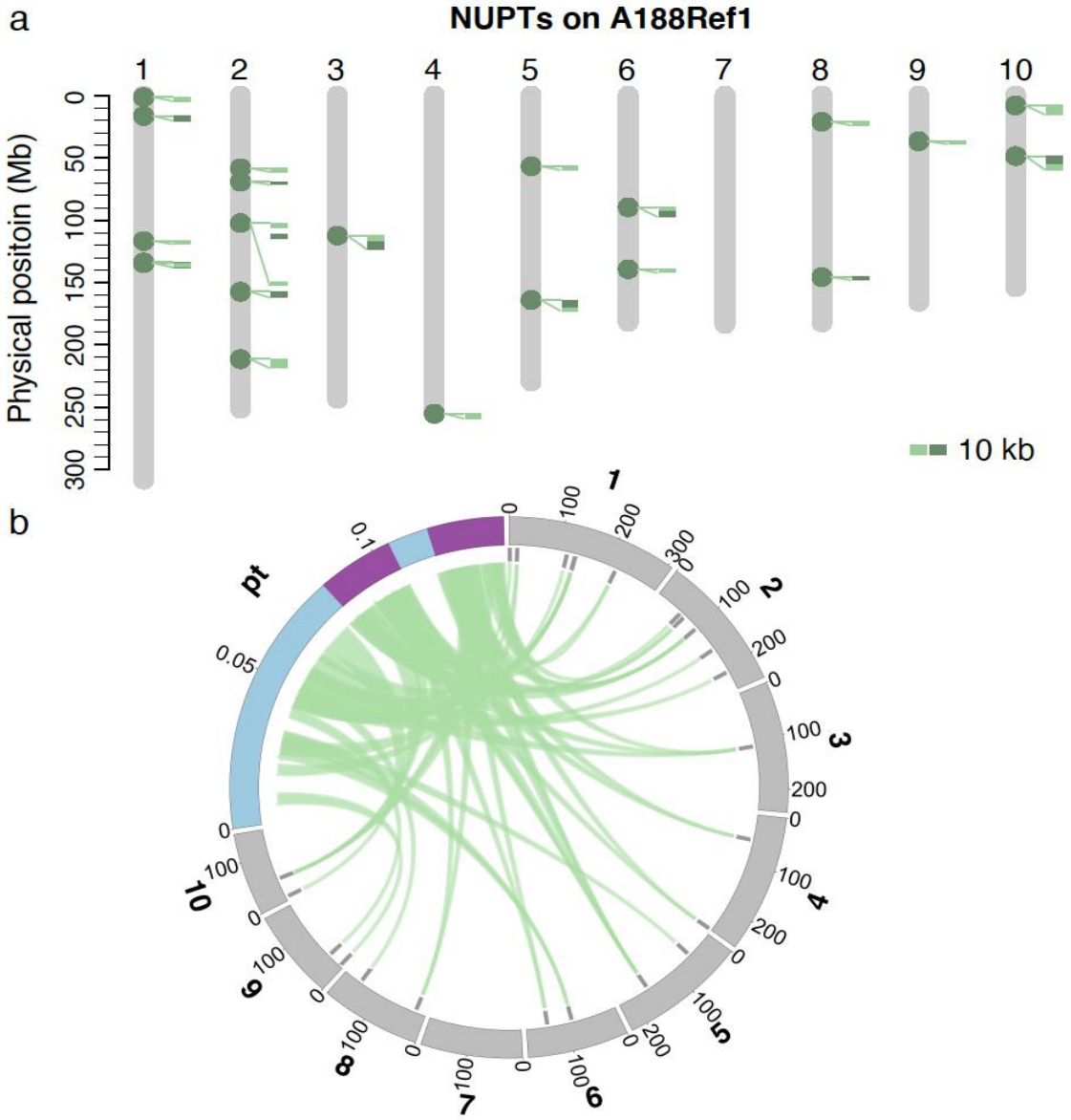

Figure B.6 Nuclear integration of chloroplast DNA. a). NUPT sequence on 10 chromosomes of A188Ref1. Each dot on chromosomes designates a potential NUPT integration. Close-up alignments with the chloroplast genome are shown along NUPTs. Each alignment requires at least 3 Kb match and 95% identity. **b**). Circos plot of alignments between the chloroplast (pt) genome and ten chromosomes. Purple highlighted the large duplicated regions on pt. Gray bars locate NUPT positions. Note that the chromosomal scale is different from the pt scale. Numbers on the track are in Mb.

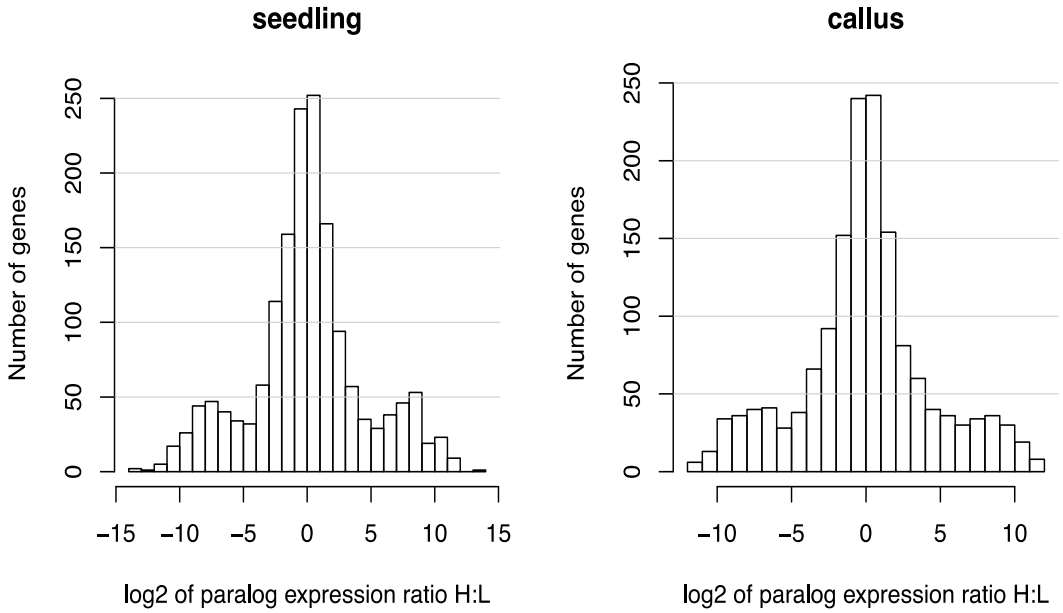

Figure B.7 Expression comparison of paralogs in high- and low-recombination regions. Pairs of paralogs of which one is located at a high-recombination region (H) and the other is located at a low-recombination region (L) were compared for their expression. Histograms of the log2 values of read counts ratio of H to L were plotted. Relatively symmetric distributions between positive and negative log2 values indicated that the genomic context of the gene location was not a major driver for gene expression.

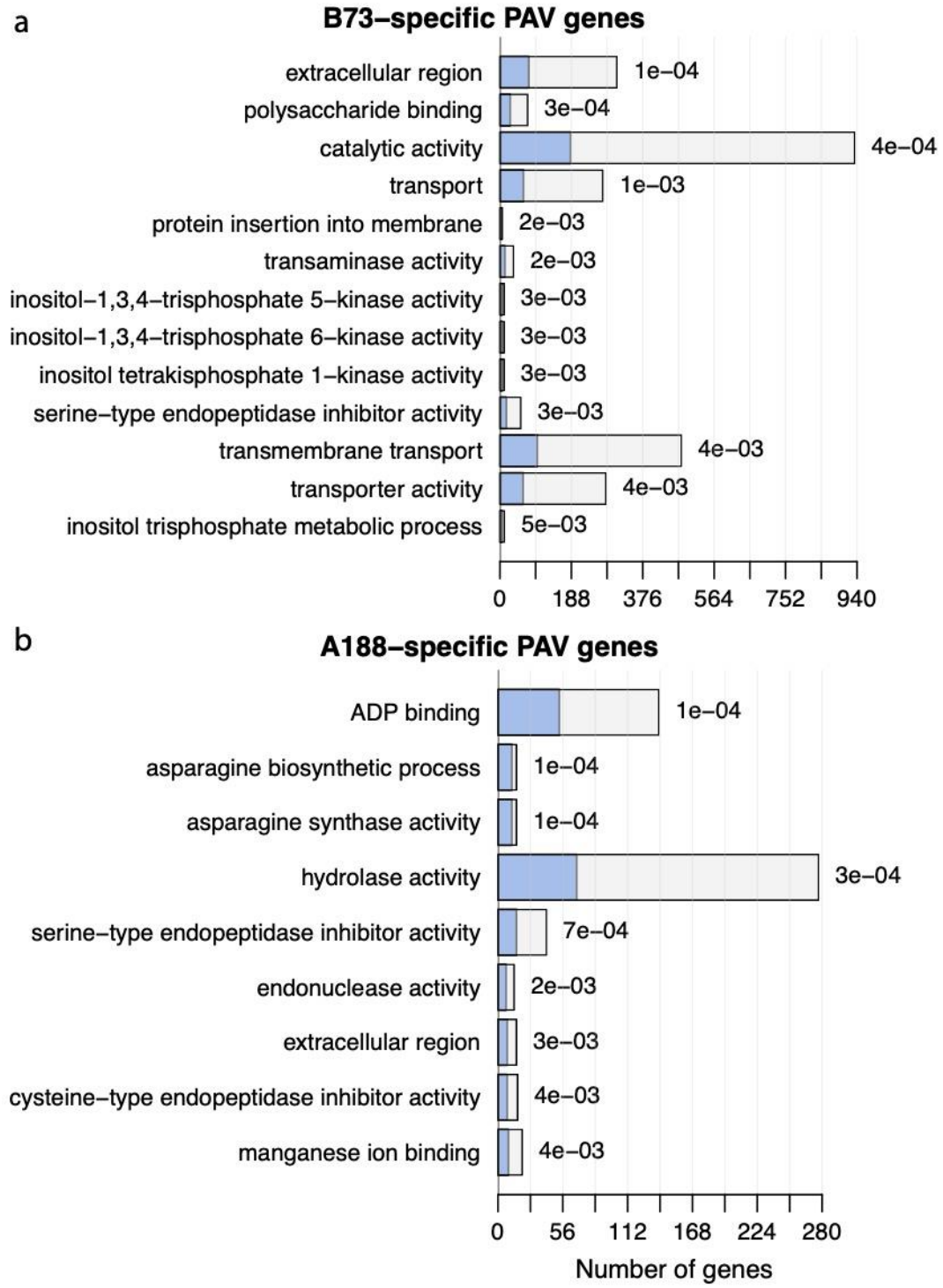

Figure B.8 GO enrichments of PAV/HDS genes. Enriched GO terms in B73-specific PAV/HDS genes (**a**) and in A188-specific PAV/HDS genes (**b**). In each barplot, a blue bar stands for the number genes in the PAV/HDS gene set and the whole bar (blue and empty) stands for the total number of genes of the associated GO term. P-values were labeled on the top of each bar. Only

the GO terms with the p-value smaller than 0.005 and containing at least five PAV genes were plotted.

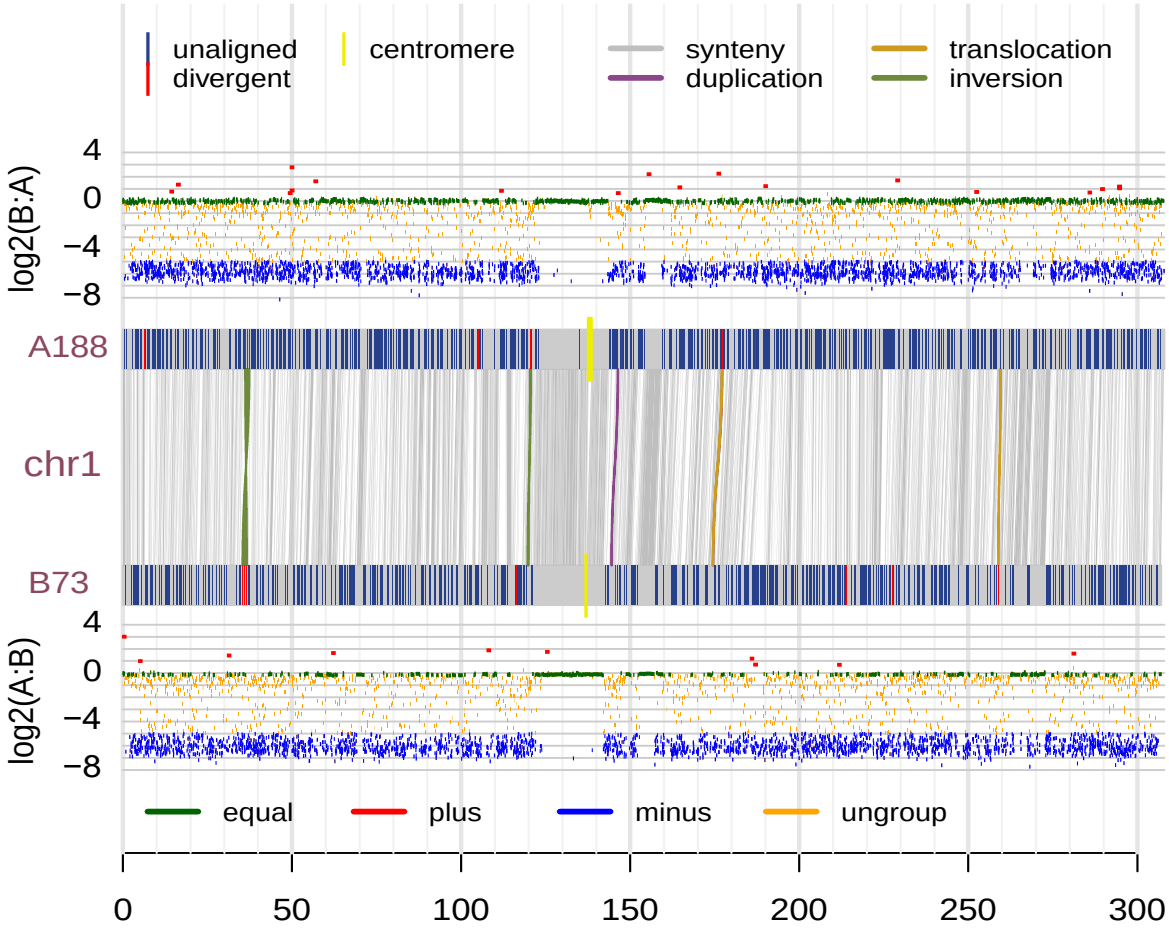

Figure B.9 SyRI and CGRD results on chromosome 1. CGRD results using A188Ref1 and B73Ref4 as the reference genomes were plotted on the top and bottom, respectively. Y-axis represents log2 values of ratios of read depths of B73 to A188, log2(B:A), or log2 values of ratios of read depths of A188 to B73, log2(A:B), signifying copy number variation (CNV). The SyRI result is displayed in between two CGRD results. Alignments of syntenic blocks larger than 10 Kb and alignments of other rearrangements larger than 0.5 Mb are plotted. On each A188 and B73 chromosome, segments not aligned to the other genome (unaligned), segments divergent with the other genome in a high degree (divergent), and centromeres are highlighted. The same plotting strategy was applied to chromosome 2, 3, 5, 6, 7, 8, 9, and 10.

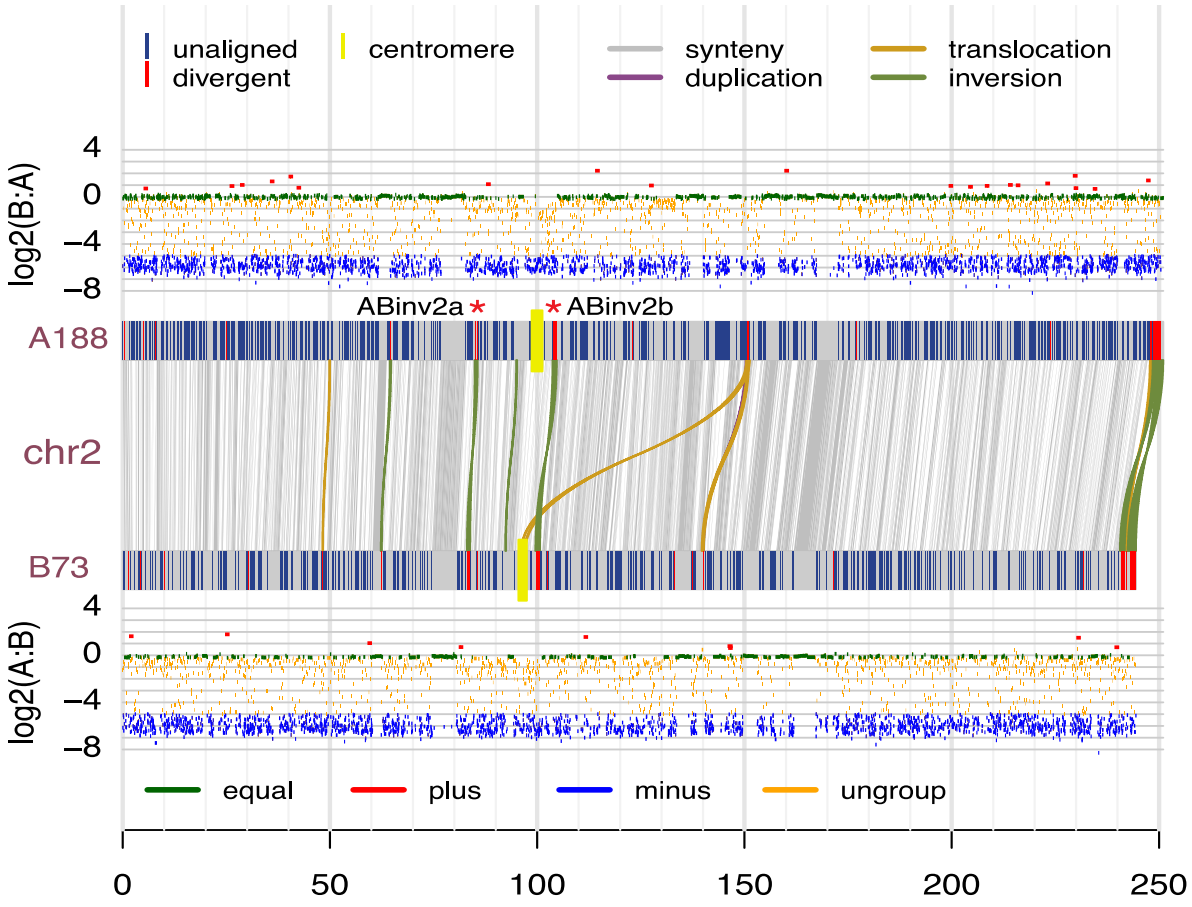

Figure B.10 SyRI and CGRD results on chromosome 2. The red * labels a well-evidenced inversion.

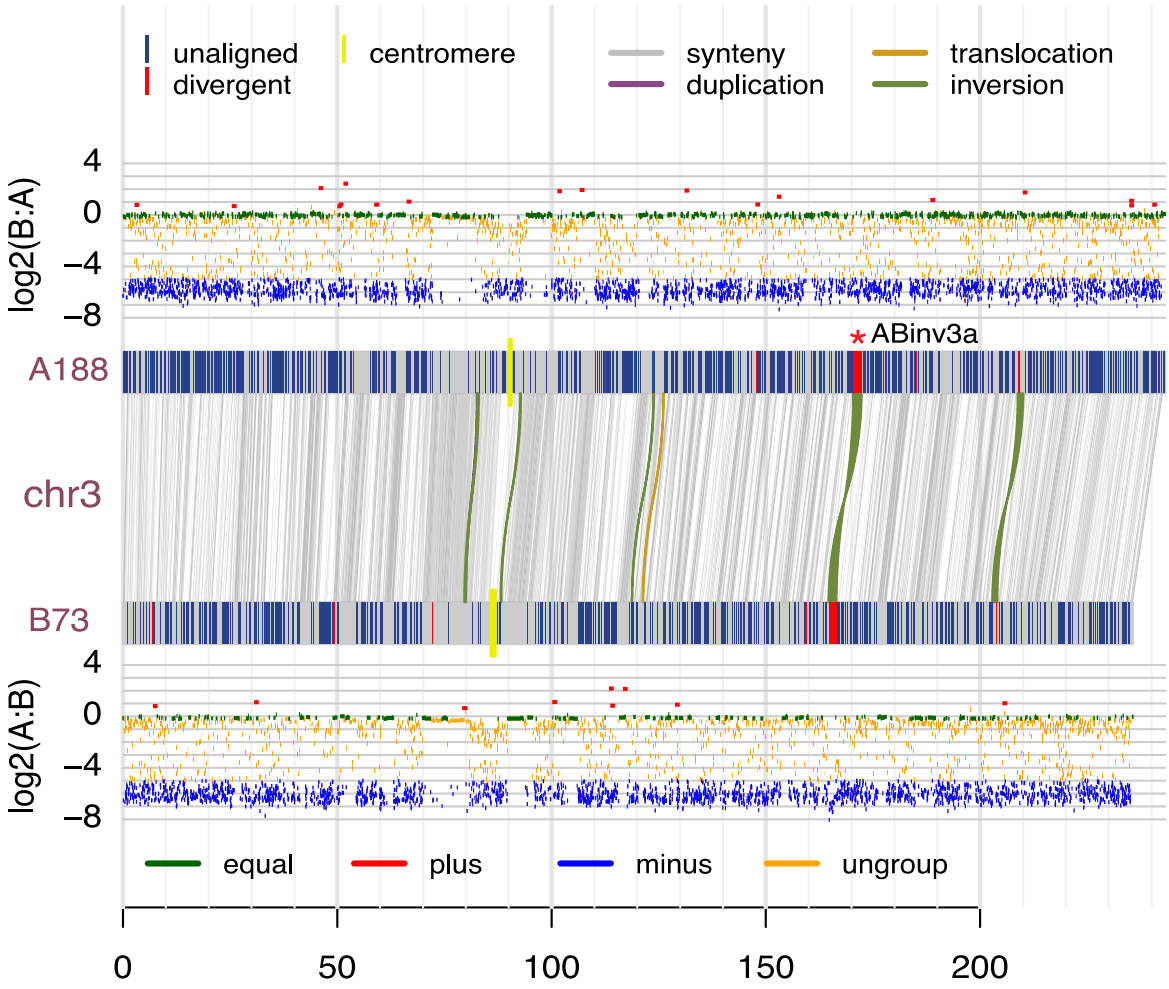

Figure B.11 SyRI and CGRD results on chromosome 3. The red * labels a well-evidenced inversion.

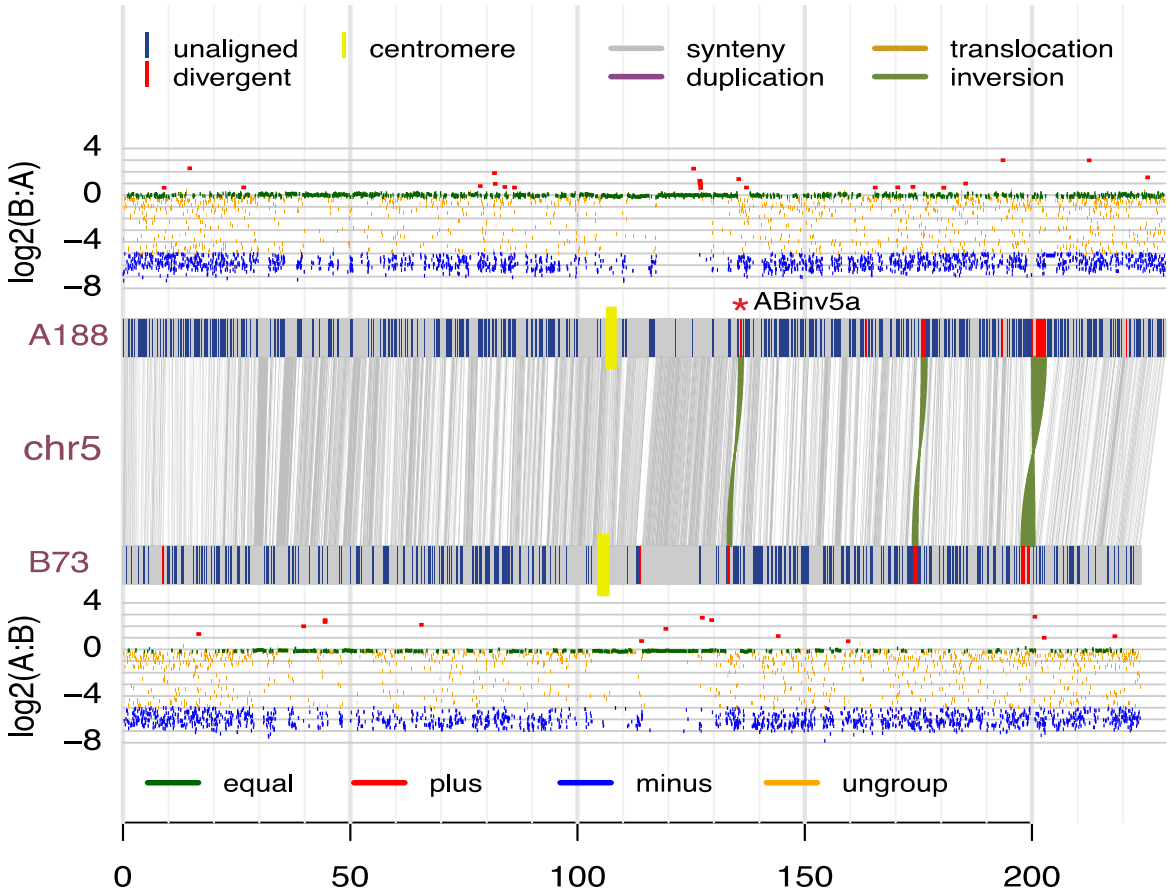

Figure B.12 SyRI and CGRD results on chromosome 5. The red * labels a well-evidenced inversion.

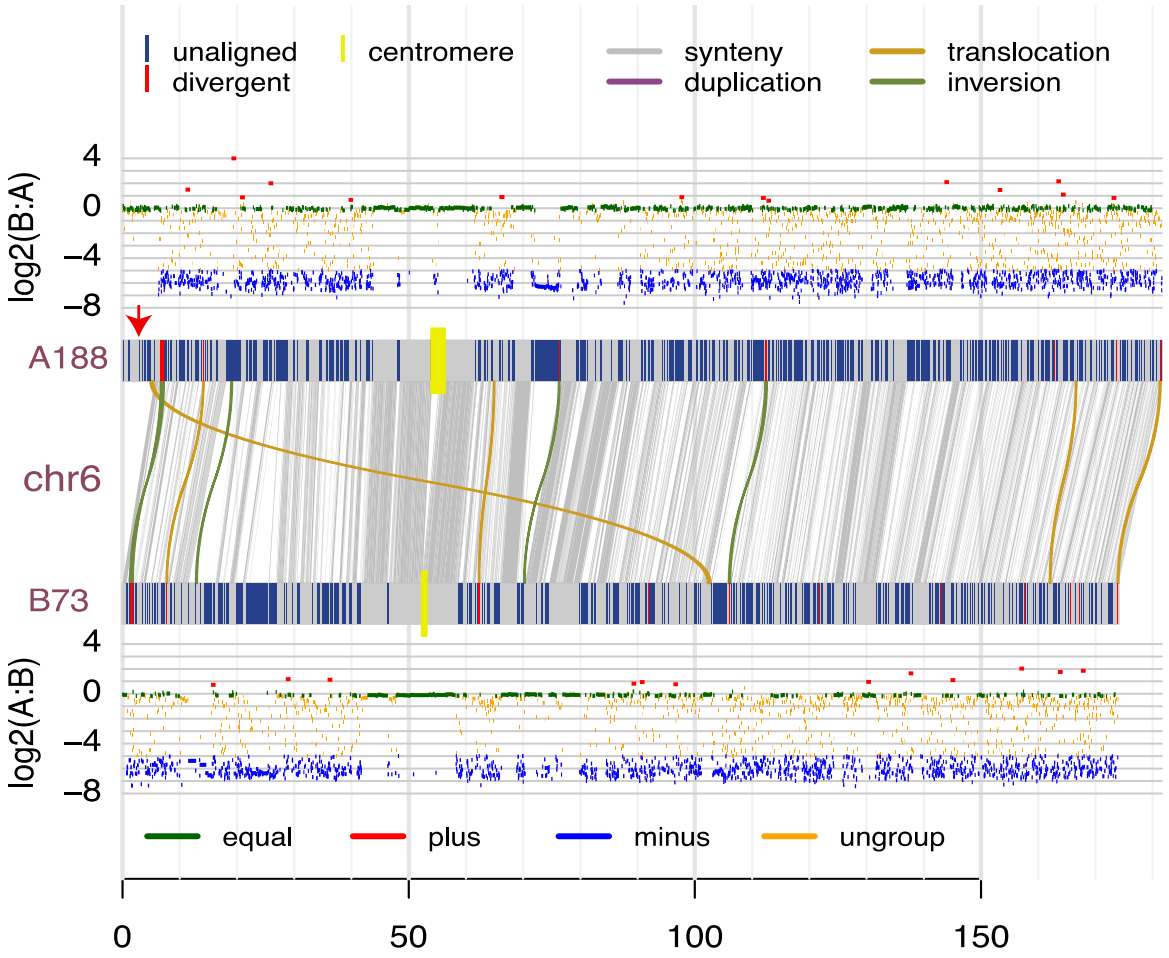

Figure B.13 SyRI and CGRD results on chromosome 6. The arrow points at a relative conserved region (equal) between the two genomes. However, the region is missed in the B73Ref4. The newly assembled B73Ref5 has the region at the beginning of chromosome 6.

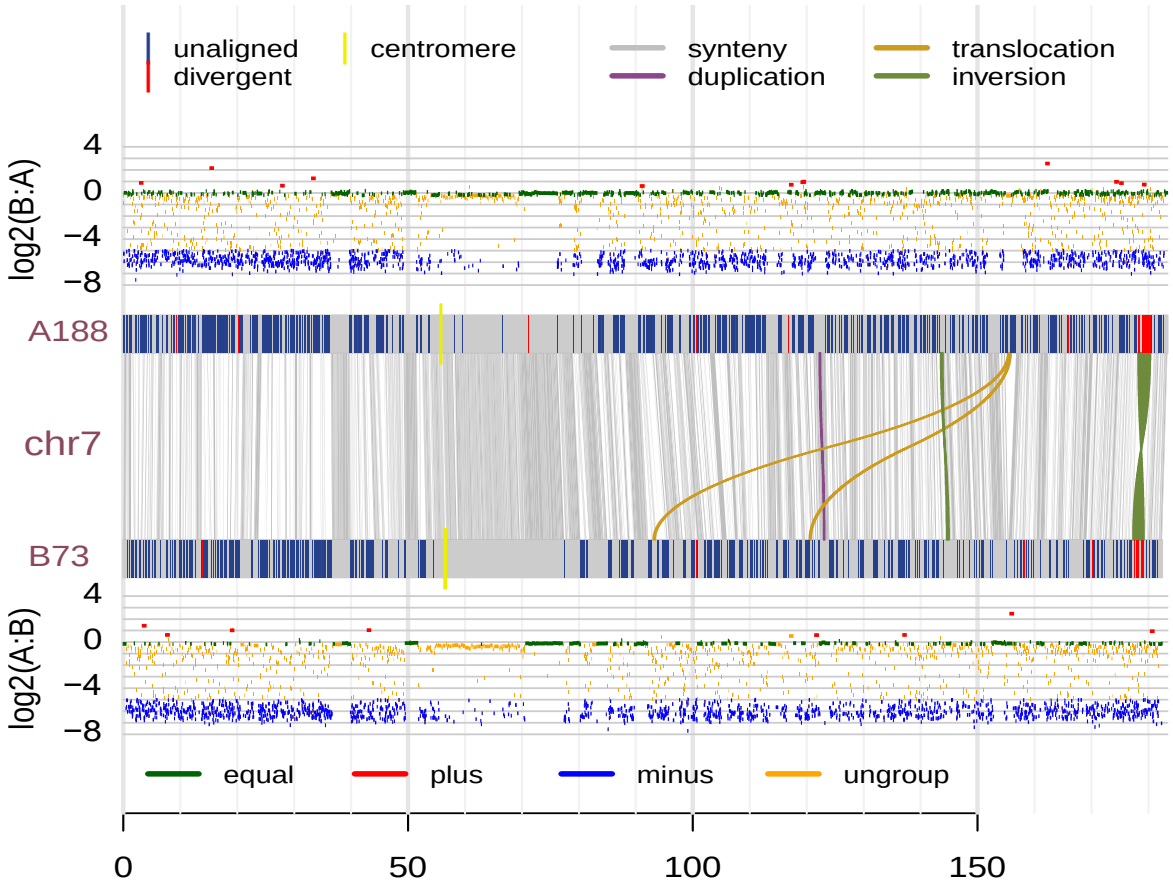

Figure B.14 SyRI and CGRD results on chromosome 7.

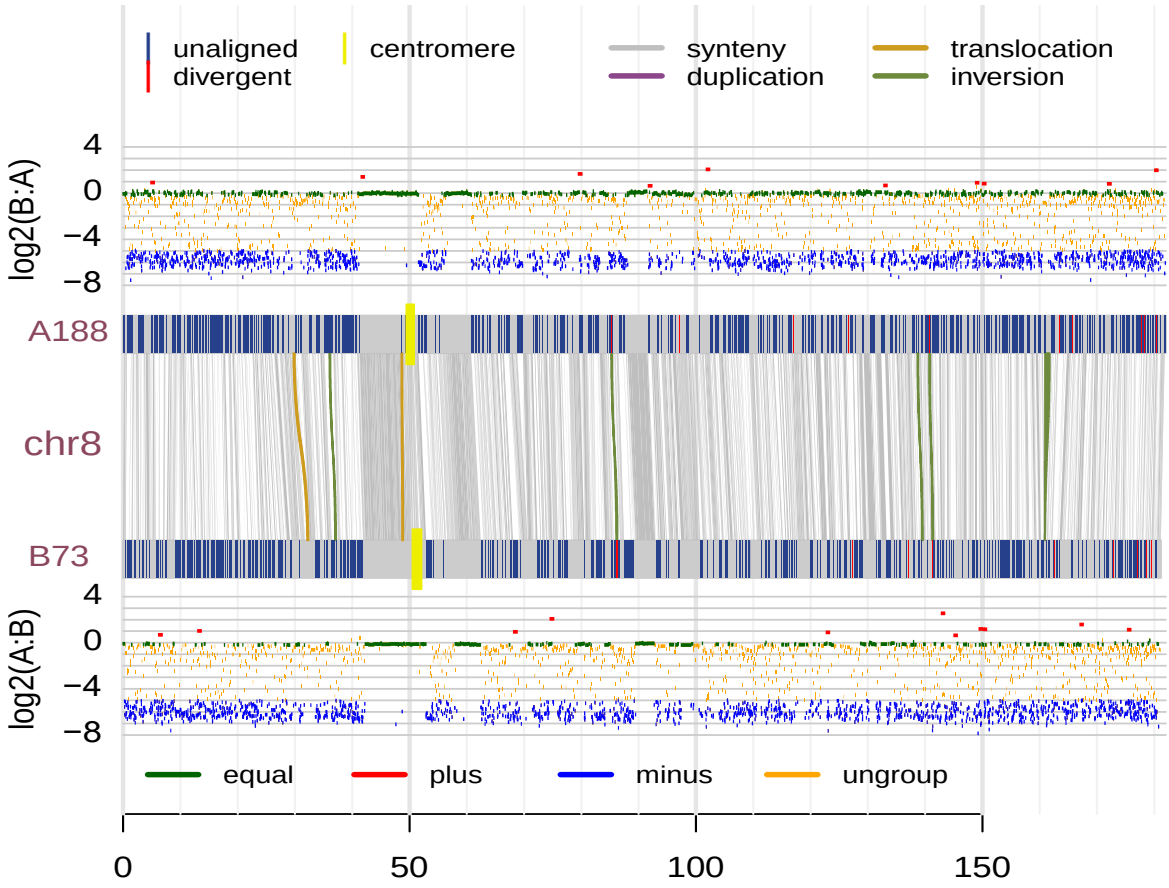

Figure B.15 SyRI and CGRD results on chromosome 8.

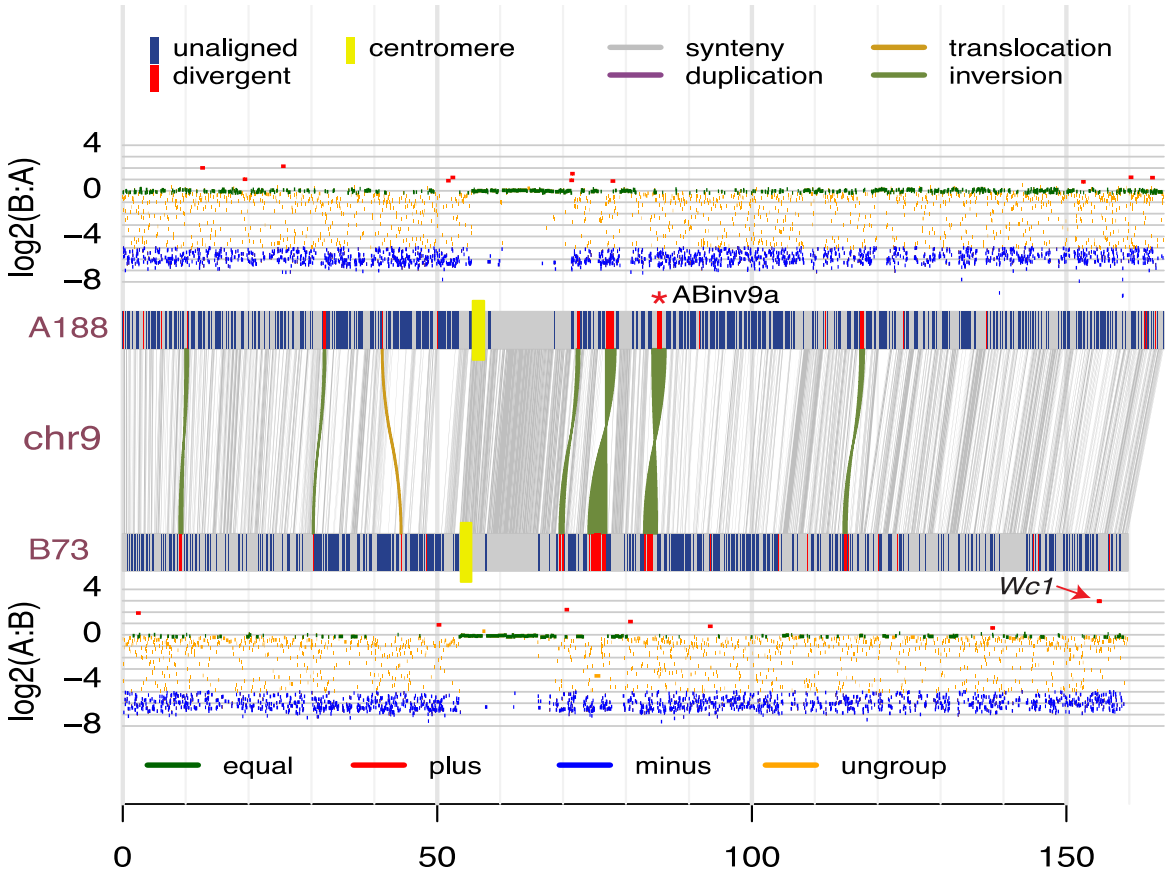

Figure B.16 SyRI and CGRD results on chromosome 9. The red * labels a well-evidenced inversion. The arrow points at the B73 *Wc1* region showing A188plus, which A188 has higher copy number as compared to B73.

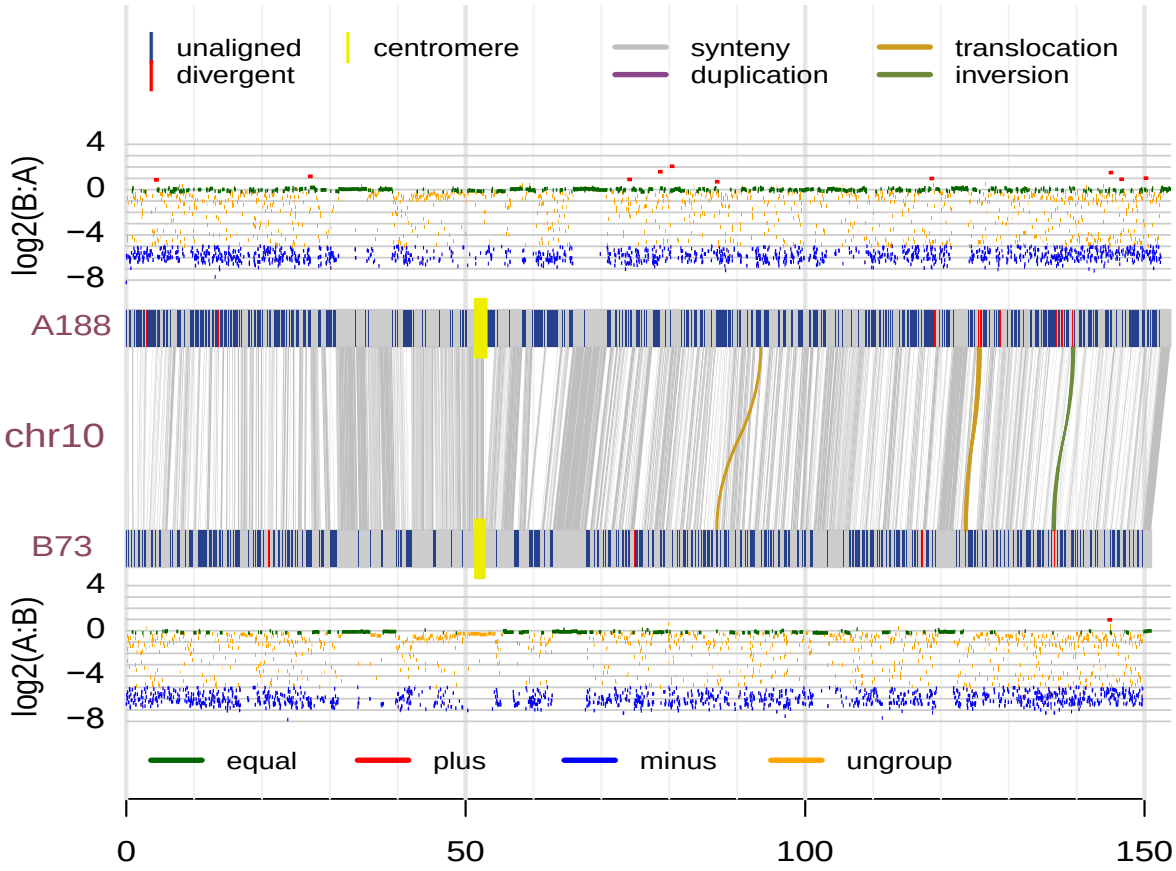

Figure B.17 SyRI and CGRD results on chromosome 10.

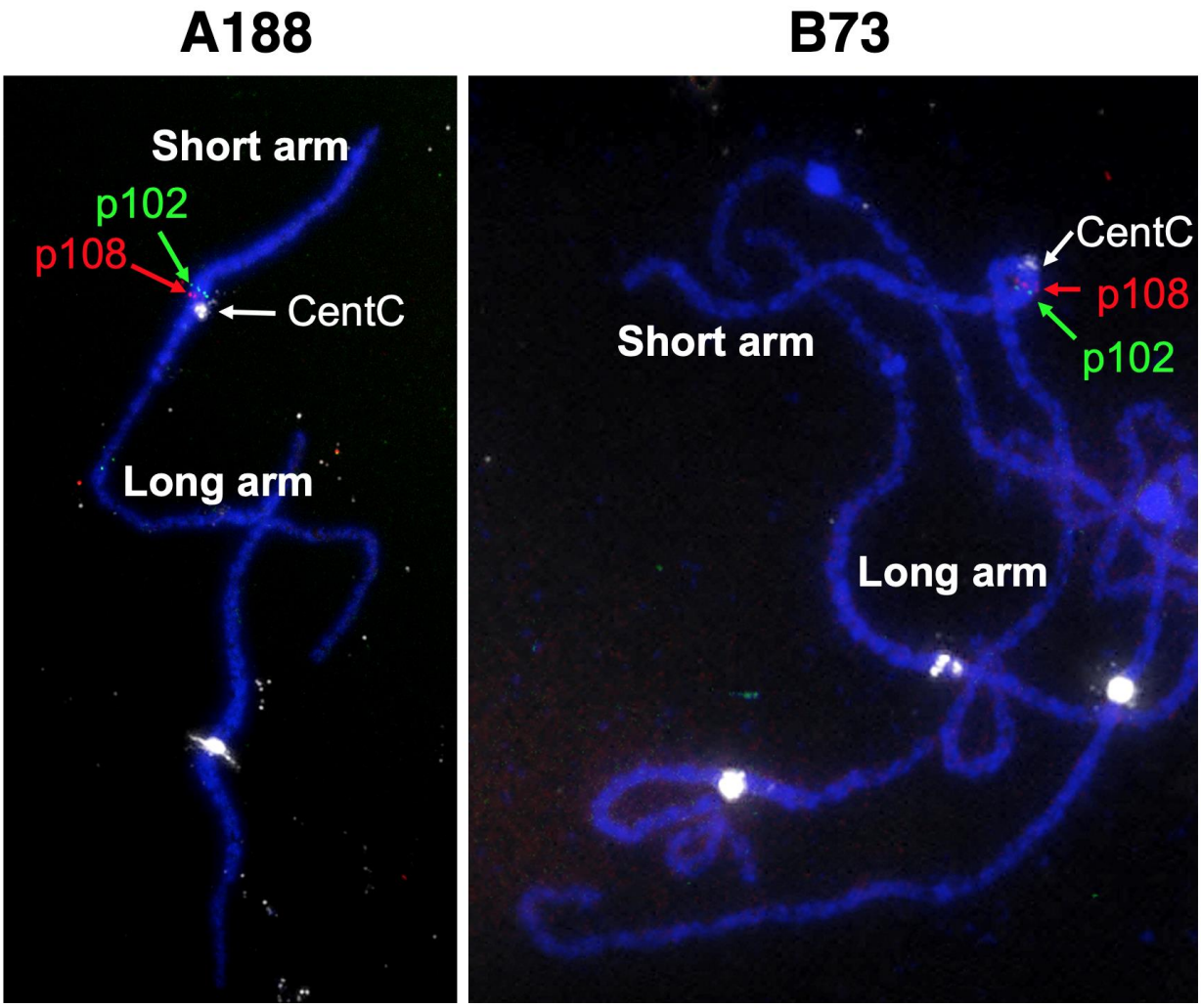

Figure B.18 FISH on an inversion candidate. Two probes (p102 and p108) were designed on the potential inversion and labeled with green and red fluroscent colors. The CentC probe was used to locate the centromere. The result indicated that both A188 and B73 had the same order of p102 and p108, which did not support the inversion.

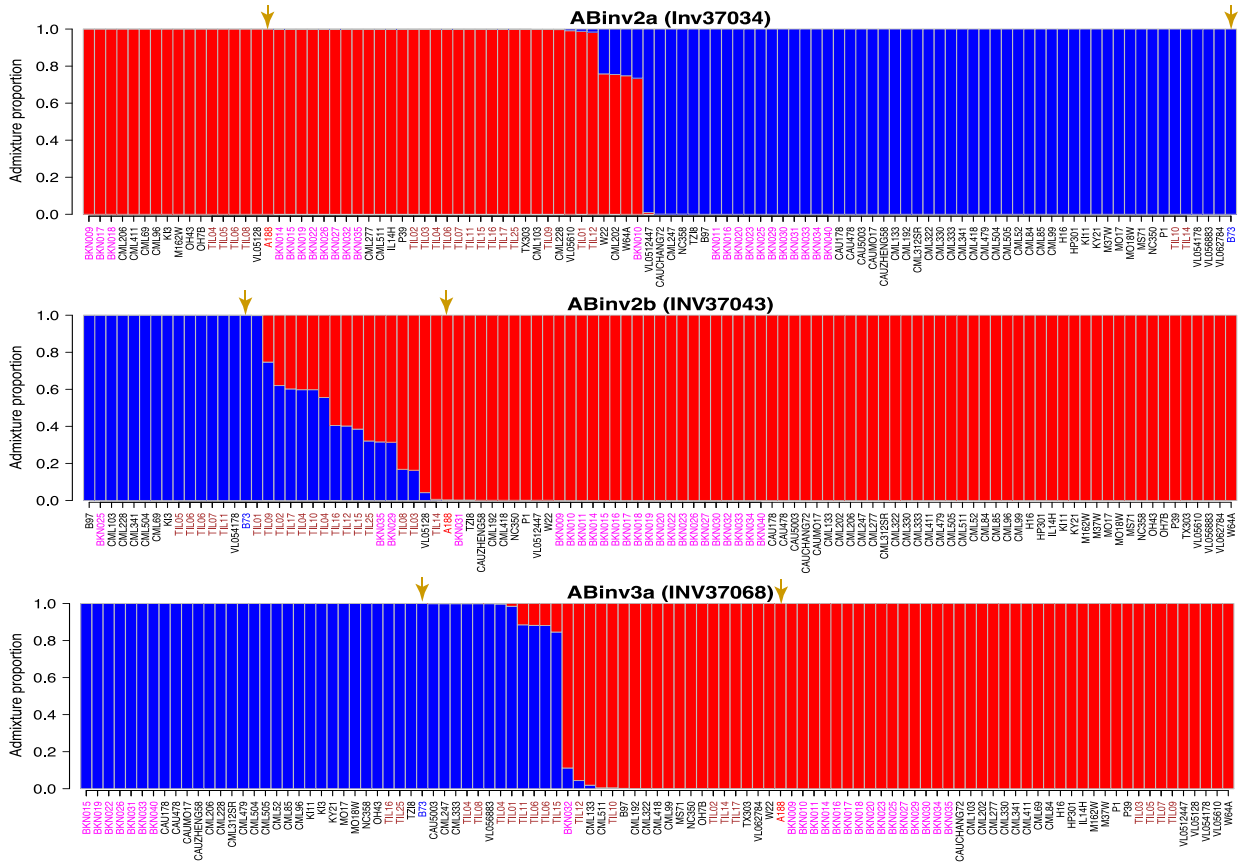

Figure B.19 Structure analysis of three inversions in the maize Hapmap2 population. The x-axis represents the maize lines. The y-axis represents the admixture proportion of two sub-populations for each line. Arrows point at A188 and B73. The maize wild ancestors, teosinte lines, are highlighted in brown, and landrace lines are highlighted in magenta.

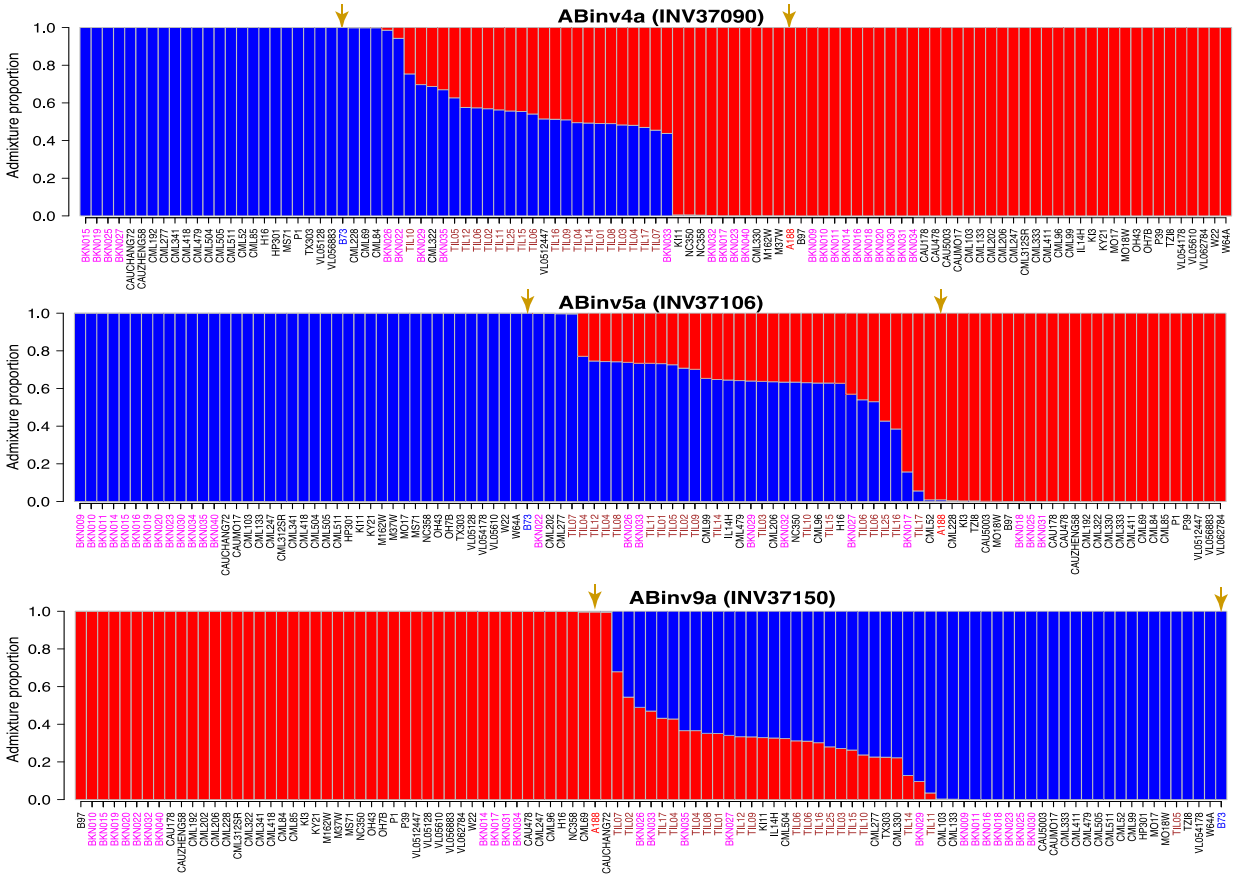

Figure B.20 Structure analysis of other three inversions in the maize Hapmap2 population - II. The x-axis represents the maize lines. The y-axis represents the admixture proportion of two sub-populations for each line. Arrows point at A188 and B73. The maize wild ancestors, teosinte lines, are highlighted in brown, and landrace lines are highlighted in magenta.

a

b

Genotype combination of DHs			DH count of color codes from 1-6					
QTL_chr2	QTL_chr6	QTL_chr9	1	2	3	4	5	6
A	A	A	13	1	0	0	0	0
B	A	A	7	6	5	0	0	0
A	A	B	3	0	4	0	0	0
B	A	B	5	1	5	0	1	1
A	B	A	5	4	4	0	1	0
B	B	A	0	3	11	6	2	1
A	B	B	0	0	1	0	4	0
B	B	B	0	2	1	1	13	8

Figure B.21 Phenotypic and genotypic data of DH lines. a) A standard of kernel colors coded from 1-6. **b**) Counts of DH lines with kernel colors matching to standard codes for each genotype combination of three kernel color QTLs.

```
A188 1   MAIILVRAASPGLSAADSISHQGTLQCSTLLKTKRPAARRWMPCSLLGLHPWEAGRPSPA 60
B73  1   MAIILVRAASPGLSAADSISHQGTLQCSTLLKTKRPAARRWMPCSLLGLHPWEAGRPSPA 60

A188 61  VYSSLAVNPAGEAVVSSEQKVYDVVLKQAALLKRQLRTPVLDARPQDMDMPRNGLKEAYD 120
B73  61  VYSSLAVNPAGEAVVSSEQKVYDVVLKQAALLKRQLRTPVLDARPQDMDMPRNGLKEAYD 120

A188 121 RCGEICEEYAKTFYLGTMLMTEERRRAIWAIYVWCRRTDELVDGPNANYITPTALDRWEK 180
B73  121 RCGEICEEYAKTFYLGTMLMTEERRRAIWAIYVWCRRTDELVDGPNANYITPTALDRWEK 180

A188 181 RLEDLFTGRPYDMLDAALSDTISRFPIDIQPFRDMIEGMRSDLRKTRYNNFDELYMYCYY 240
B73  181 RLEDLFTGRPYDMLDAALSDTISRFPIDIQPFRDMIEGMRSDLRKTRYNNFDELYMYCYY 240

A188 241 VAGTVGLMSVPVMGIASESKATTESVYSAALALGIANQLTNILRDVGEDARRGRIYLPQD 300
B73  241 VAGTVGLMSVPVMGIATESKATTESVYSAALALGIANQLTNILRDVGEDARRGRIYLPQD 300

A188 301 ELAQAGLSDEDIFKGVVTNRWRNFMKRQIKRARMFFEEAERGVTELSQASRWPVWASLLL 360
B73  301 ELAQAGLSDEDIFKGVVTNRWRNFMKRQIKRARMFFEEAERGVTELSQASRWPVWASLLL 360

A188 361 YRQILDEIEANDYNNFTKRAYVGKGKKLLALPVAYGKSLLLPCSLRNGQT 410
B73  361 YRQILDEIEANDYNNFTKRAYVGKGKKLLALPVAYGKSLLLPCSLRNGQT 410
```

Figure B.22 Alignment between Y1 protein sequences of A188 and B73. Protein sequences of the transcript Zm00056a032392*T003 (A188) and the transcript Zm00001d036345*T001 (B73) were compared. Of 410 amino acids, 409 were identical. The polymorphic site is highlighted in red.

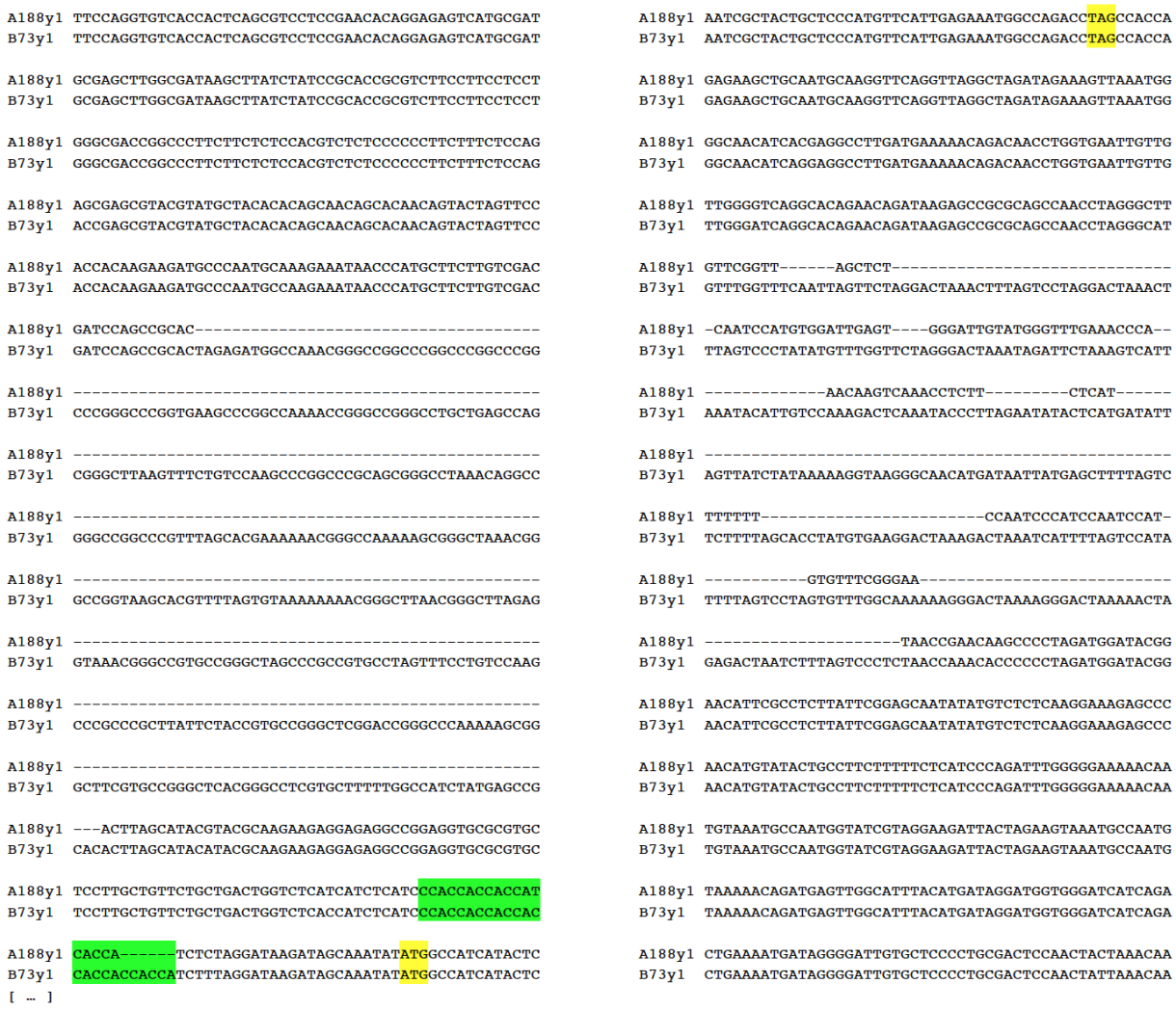

Figure B.23 Alignment of 5' and 3' flanking sequences of A188 *y1* and B73 *Y1* alleles. Translation start sites and translation termination sites are highlighted in yellow. A (CCA)n microsatellite variation at the 5' untranslated region is highlighted in green. Most gene body sequences are skipped.

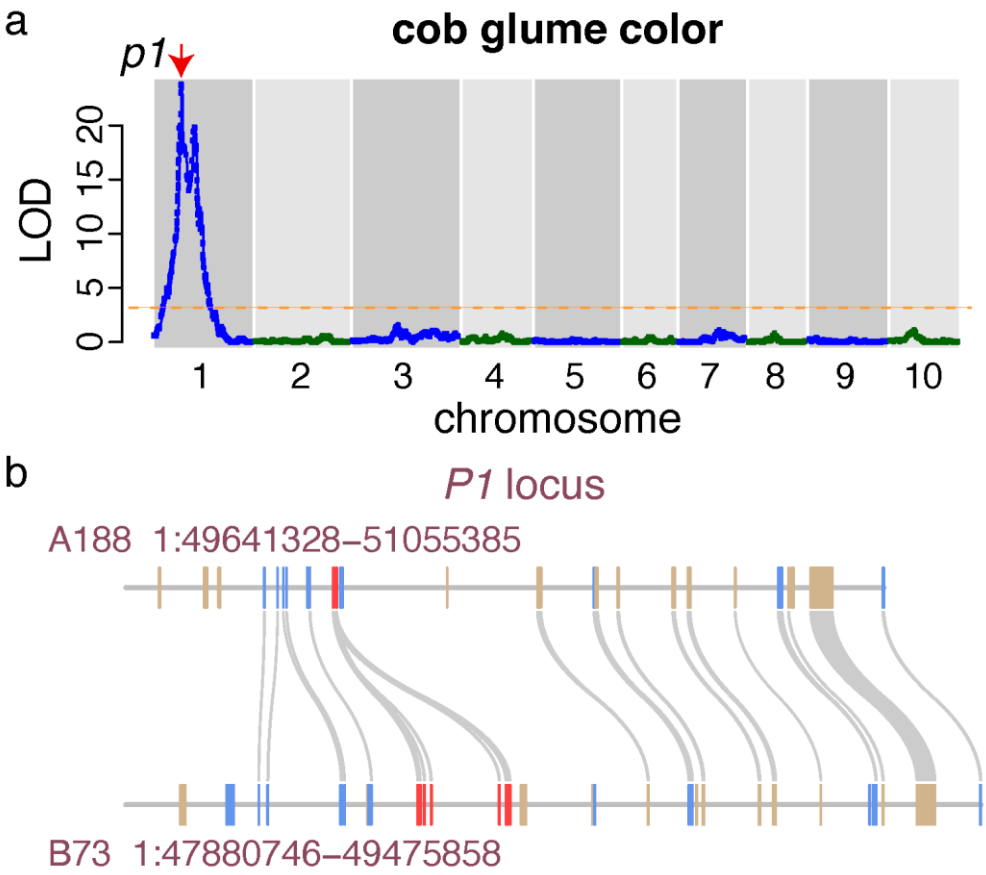

Figure B.24 Genetic analysis of cob color. a) Plot of LOD from QTL analysis versus genetic positions of markers along ten chromosomes. The orange horizontal line indicates the LOD threshold from 1,000 permutation test. The arrow points at the genetic location of the *P1* gene. **b**). Each rectangle box represents a gene with blue, tan, and red colors indicating plus, minus orientation, and *P1* homologous genes.

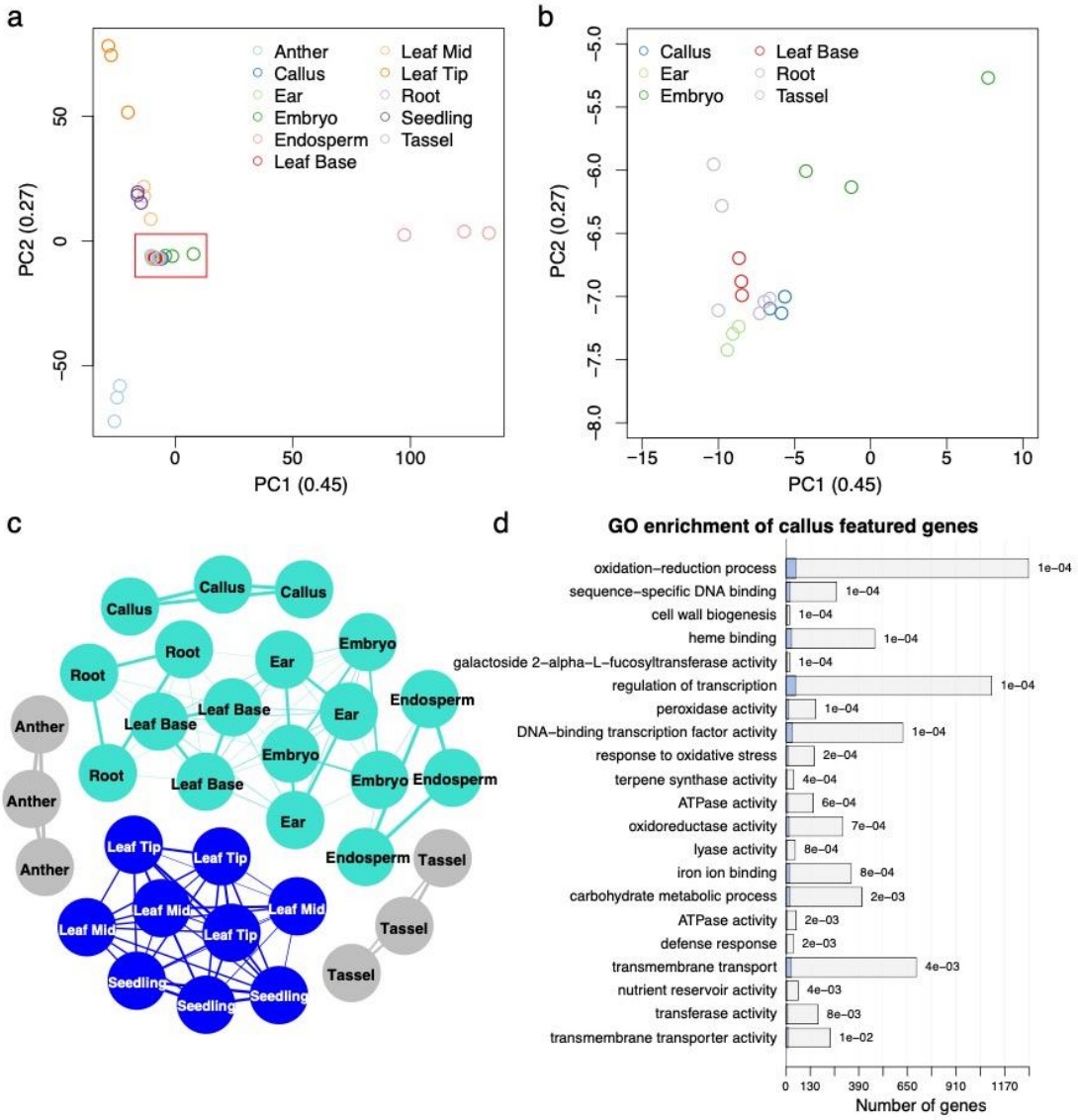

Figure B.25 sample clustering and callus-featured genes a) Principal component analysis (PCA) results of gene expressions in 11 A188 tissues. The x-axis and y-axis represent the first component (PC1) and the second component (PC2), respectively. The numbers within the parentheses stand for the proportions of the variation of gene expressions explained by either PC1 or PC2. **b**) Enlarged PCA plot of the red box in **a**. **c**) The network of 33 RNA-Seq samples from 11 tissue types constructed based on their gene expression. Two major clusters were identified. One cluster (turquoise) includes callus, root, leaf base, ear, embryo, and endosperm; the other cluster (blue) includes leaf tip, leaf middle, and seedling. The leaf base, middle, and tip are three parts from base to tip from the same leaf. **d**) GO enrichment of callus featured genes. In each barplot, a blue bar stands for the number callus featured genes and the whole bar (blue and empty) stands for the total number of genes of the associated GO term. P-values were labeled on the top of each bar. Only the GO terms with the p-value smaller than 0.01 and containing at least five callus featured genes were plotted.

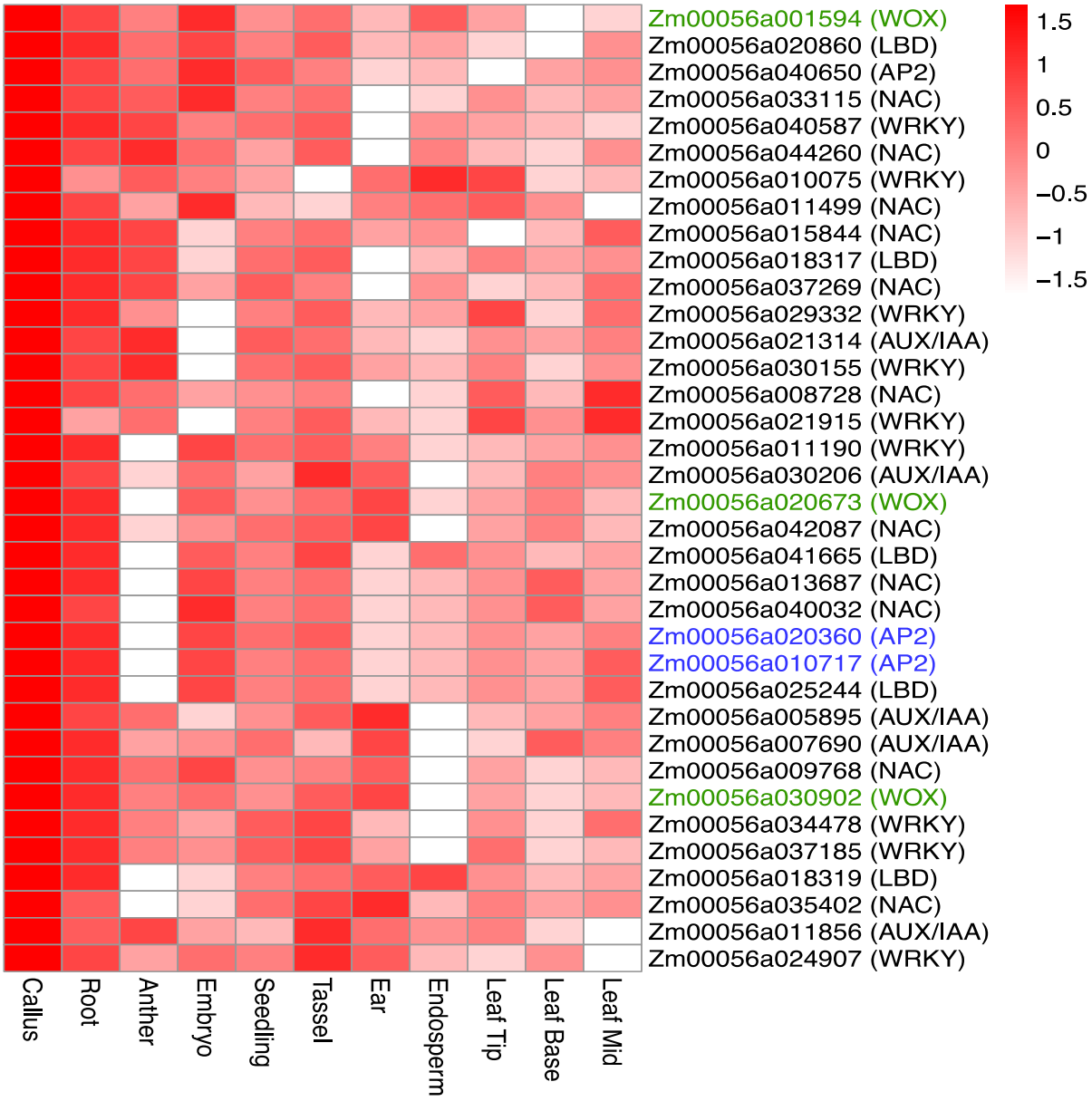

Figure B.26 Heatmap plots of expression of callus-featured TF genes. The row and column represent the callus-featured TFs and tissues, respectively. The values of gene-wise quantile-quantile normalized (qqnorm) gene expressions are color-coded. The qqnorm implemented by an R package "qqnorm" normalized gene expressions to a Gaussian distribution. Genes homologous to *Baby boom* and *Wuschel2* are colored in blue and green, respectively.

Supplemental Tables

Table B.1 Phenotype comparison between A188 and B73

No.	Trait	A188	B73	Trait description	Reference
1	Tissue Culturability	high	low	A188: white, compact and nodulated callus; B73: brown, loose and non-embryogenic callus	this study
2	Starch	65.813	69.573	Starch BLUPs* of kernel across environments (%)	1
3	Protein	14.572	13.058	Protein BLUPs of kernel across environments (%)	1
4	Oil	4.624	3.611	Oil BLUPs of kernel across environments (%)	1
5	DTS	67.996	75.789	DTS BLUPs across environments; days from planting date to silking date	2
6	DTA	66.173	75.552	DTA BLUPs across environments; days from planting date to tasseling date	2
7	ASI	0.775	0.636	ASI BLUPs across environments; anthesis-silking interval	2
8	PH	126.094	178.731	PH BLUPs across environments; distance from soil surface to the base of flag leaf (cm)	2
9	EH	45.442	88.893	EH BLUPs across environments; distance from soil surface to the highest ear-bearing node (cm)	2
10	Tassel Primary Branches	18.5	7.79	Number of primary branches; data collected from 2006 Illinois field	3
11	Upper Leaf Angle	45	81.5	Angle between the stalk and the midrib of leaf immediately below the flag leaf; vertical = 90 and horizontal = 0 degrees; data collected from 2006 Illinois field	3
12	CobDiameter	25	30.57	Diameter of cob at halfway point of length; data collected from 2006 Illinois field	3
13	CobWeight	12	23.92	Mass of cob only; data collected from 2006 Illinois field	3

14	EarDiameter	43	40.69	Diameter of ear at 1/5 length point with kernels on it or the widest point; data collected from 2006 Illinois field	3
15	EarRowNumber	13	16.91	Number of rows around circumference; data collected from 2006 Illinois field	3
16	EarRankNumber	13	24.23	Number of rows along length; data collected from 2006 Illinois field	3
17	Goss's Wilt Lesion Length	19.6	9.4	Average leaf lesion inoculated with bacterium Clavibacter michiganensis subsp. nebraskensis (Cmn)	4
18	SLB	242	192.5	Rank of resistance to fungal disease Southern leaf blight; rank 1= most resistant, rank 253=most susceptible , and ties in rank were averaged	5
19	GLS	249	196	Rank of resistance to fungal disease Gray leaf spot; rank 1= most resistant, rank 253=most susceptible, and ties in rank were averaged	5
20	NLB	251.5	202	Rank of resistance to fungal disease Northern leaf blight; rank 1= most resistant, rank 253=most susceptible, and ties in rank were averaged	5
21	MDR	248	201	Rank of resistance to multiple diseases; rank 1= most resistant, rank 253=most susceptible, and ties in rank were averaged	5

Reference

1. JP Cook et al. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association Panels. Plant Physiology, 158(2), 824-834.
2. JA Peiffer et al. (2014). The genetic architecture of maize height. Genetics, 196(4), 1337-1356.
3. panzea.com
4. Y Hu et al. (2018). Analysis of Extreme Phenotype Bulk Copy Number Variation (XP-CNV) Identified the Association of rp1 with Resistance to Goss's Wilt of Maize. Frontiers in Plant Science, 9.
5. R. J. Wisser et al. (2011). Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. Proc. Natl. Acad. Sci. U. S. A. 108: 7339–7344.
* BLUPs stands for best linear unbiased predictors

Table B.2 List of tissue types for RNA-Seq

Order	SampleID	Tissue type	Description
R01	root	root	Root of 10-day-old seedling
R02	seedling	seedling	Above-ground of 10-day-old seedling
R03	leaf11v12base	leaf base	Base of the 11th leaf at V12
R04	leaf11v12mid	leaf mid	Middle of the 11th leaf at V12
R05	leaf11v12tip	leaf tip	Tip of the 11th leaf at V12
R06	tasselv18	tassel	Meiotic tassel at V18
R07	earv18	ear	Immature ear at V18
R08	antherR1	anther	Anthers at R1
R09	endospermDAP16	endosperm	Endosperm 16 days after pollination
R10	embryoDAP16	embryo	Embryo 16 days after pollination
R12	callus	callus	Callus 39 days after tissue culture on CIM

Table B.3 List of primers used for FISH probes and qRT-PCR

No.	Primer name	Primer (5'--3')	PCR size (bp)	Notes
1	NUMT_1F	GCCAGTGTGCTGGAATTCCACTCCGTTCCCGACATAGT	8,950	NUMT FISH probe; taget region of A188 Ref1 8:30785164-30794113, Fusion clone
2	NUMT_2R	TGGATATCTGCAGAATTCAGGGCATTGAACGAGAGAGA		
3	PME45501_F	GTCGGTGCCACCATCATCAT	484	PME FISH probe
4	PME45501_R	CTAGCGTTGTTGCAGCCCTA		
5	cent4f2	CCCTAAGCACTAAACGCTGA	161	Cent4 FISH probe specifically targeting on the centromere of chromosome 4
6	cent4r2	GGTTTGATGTTTAGGTTGGACA		
7	Ach4_102_4F	GCCAGTGTGCTGGAATTCTAACCAAACCCGGTTCCAGTTC	6125	p102 FISH probe; target region of A188Ref1 4:102423375-102429499
8	Ach4_102_4R	TGGATATCTGCAGAATTCGCAGTGGAAATGAGAATGATCGG		
9	Ach4_108_21F	GCCAGTGTGCTGGAATTCTCCCGCTTTGTTTGCTTGAATT	4079	p108 FISH probe part 1; target region of A188Ref1 4:106308779-106312857
10	Ach4_108_21R	TGGATATCTGCAGAATTCGTACGCTAGTACCCACACATCC		
11	Ach4_108_23F	GCCAGTGTGCTGGAATTCACTCATGAACCGCAATCAATGT	2758	p108 FISH probe part 2; target region of A188Ref1 4:106331369-106334126
12	Ach4_108_23R	TGGATATCTGCAGAATTCAAACAGTAGCGGTCTATGGAGC		
13	act1_qrt_2F	GGCGAACAACTGGTATTGTGAT	98	Primers for actin1 amplification in qRT-PCR
14	act1_qrt_2R	CAAACGGAGAATAGCATGAGGA		
15	ccd1_qrt_5F	CGGGAATGAGCTGTACGAGAT	123	Primers for ccd1 amplification in qRT-PCR
16	ccd1_qrt_5R	CGCTGCTTTCTGCCAGTATAG		
17	y1_qrt_4F	CATAGACATTCAGCCATTCAG	119	Primers for y1 amplification in qRT-PCR
18	y1_qrt_4R	CGACAGTTCCAGCAACATAGTA		

Table B.4 Summary of contigs in the A188Ref1 assembly

Item	Value
total contig number*	1,854
total contig length (bp)	2,235,724,030
mean contig length (bp)	1,205,892
median contig length (bp)	126,948
longest contig (bp)	50,817,798
shortest contig (bp)	16,204
N50 (bp)	5,994,828
L50	89
overall GC	0.47
mean contig GC	0.47
median contig GC	0.47
max contig GC	0.65
min contig GC	0.35

* some contigs were split during hybrid scaffolding with Bionano data

Table B.5 Summary of corrected errors during four polishing steps

Step	Total	Insert	Substitution	Deletion	Replace*
Nanopolish 1	46,552,455	31,155,404	11,881,754	3,207,935	153,681
Nanopolish 2	21,342,514	13,139,856	5,669,755	2,346,111	93,396
Pilon 1	40,973,979	9,000,284	18,569,312	13,404,383	-
Pilon 2	1,625,385	279,539	981,022	364,824	-

* multiple-base replacements

Table B.6 Summary of SyRI structural annotation

Variant_type	Count	Minimum length (Kb)	Length_B73 (bp)	Length_A188 (bp)
Syntenic regions	37,007	1	1,122,192,351	1,126,840,410
Inversions	148	1	37,160,763	35,886,525
Inversions	70	10	36,798,027	35,590,296
Inversions	17	500	29,916,961	28,690,148
Translocations	36,450	1	191,596,457	191,059,424
Translocations	2,302	10	42,653,436	42,322,095

Duplications (B73)	87,423	1	401,687,099	-
Duplications (B73)	4,083	10	70,551,873	-
Duplications (A188)	62,439	1	-	273,374,570
Duplications (B73)	2,333	10	-	37,520,420
Not aligned (B73)	86,214	-	441,924,901	-
Not aligned (A188)	84,920	-	-	543,755,237

Table B.7 List of large inversion candidates

ID	InvEvent[1]	B73chr	B73start	B73end	A188chr	A188start	A188end	cM_Mb[2]	INVvsB73Ref5[3]	LDmean[4]
ABinv2a	INV37034	2	83045872	83793031	2	84940375	85620135	0.0	yes	0.65
ABinv2b	INV37043	2	99579131	100658515	2	103715498	104678539	0.0	yes	0.79
ABinv3a	INV37068	3	164624821	166507098	3	170367093	172331905	0.3	yes	0.56
ABinv4a	INV37090	4	151294649	152973827	4	158720016	160325831	0.0	yes	0.68
ABinv5a	INV37106	5	132994394	133825709	5	135501832	136382710	0.0	yes	0.58
ABinv9a	INV37150	9	82861956	84919434	9	84173473	86292293	0.2	yes	0.61
-	INV37153	9	114548098	115075955	9	117234960	117832417	0.7	yes	0.35
-	INV37110	5	197659892	200541711	5	200002779	203033469	0.0	yes	0.15
-	INV37010	1	35373095	36684531	1	36146595	37404096	0.2	no	0.09
-	INV37055	2	240606483	241718671	2	248523299	249313543	2.8	no	0.07
-	INV37056	2	242514879	244413498	2	249359723	251008438	3.0	no	0.18
-	INV37069	3	202843461	204043629	3	208776561	210063328	2.8	no	0.07
-	INV37083*	4	97812929	103900949	4	101974058	108167496	0.3	no	0.31
-	INV37092	4	165827223	166683275	4	174016207	174985139	1.8	no	0.13
-	INV37107	5	173756284	174794826	5	175733246	176806342	0.5	no	0.08
-	INV37131	7	177323892	179176779	7	178209398	180310542	9.0	no	0.04
-	INV37149	9	74020239	76892804	9	76796966	78325987	0.4	no	0.33

[1] event ID defined by SyRI
[2] recombination rate in cM per Mb
[3] whether an inversion is redetectable using B73Ref5
[4] the mean of LDs between SNPs in distance of 0.2-0.3 Mb
* confirmed to be falsely discovered inversion

Table B.8 Summary of overall methylation levels in callus and seedling tissues

Sample	Tissue	Replicate	CG	CHG	CHH
A188022	seedling	1	85.6%	72.3%	1.5%
A188023	seedling	2	84.7%	71.4%	1.4%
A188122	callus	1	89.3%	75.1%	3.1%
A188123	callus	2	89.2%	74.3%	3.2%

Table B.9 Summary of DMRs in three sequence contexts

Sequence context	CG	CHG	CHH
number of DMRs	6,927	9,631	11,275
minimum DMR length (bp)	51	51	51
maximum DMR length (bp)	6,487	4,799	2,622
mean DMR length (bp)	200	243	272
number of DMRs (callus > seedling)	5,996	4,218	11,272
minimum methylation increased in the callus	5%	5%	17%
maximum methylation increased in the callus	63%	65%	41%
number of DMRs (callus < seedling)	931	5,413	3
minimum methylation decreased in the callus	21%	16%	24%
maximum methylation decreased in the callus	60%	60%	30%

Table B.10 Results from X^2 test for the independence of DMR and gene expression

Order	Methyl	Region	hypo_ deUp[1]	hypo_ deDn[2]	hyper_ deUp[3]	hyper_ deDn[4]	chisq	pvalue
1	CG	gene body	107	89	856	912	2.45	0.1174
2	CG	5-upstream	39	19	192	182	4.49	0.0342
3	CG	3-downstream	19	14	106	131	1.44	0.2299
4	CHG	gene body	232	156	393	377	7.61	0.0058
5	CHG	5-upstream	29	15	96	86	1.98	0.1595
6	CHG	3-downstream	35	30	63	56	0	1
7	CHH	gene body	0	0	169	175	NA	NA
8	CHH	5-upstream	1	1	465	327	0	1
9	CHH	3-downstream	0	0	563	397	NA	NA

[1,2,3,4] increased methylation (hyper) and decreased methylation (hypo) in the callus relative to the seedlling; up-regulated gene expression (deUp) and down-regulated gene expression (deDn) in the callus relative to seedling.

Appendix C - Supplemental data of Chapter 3

Supplemental Figures

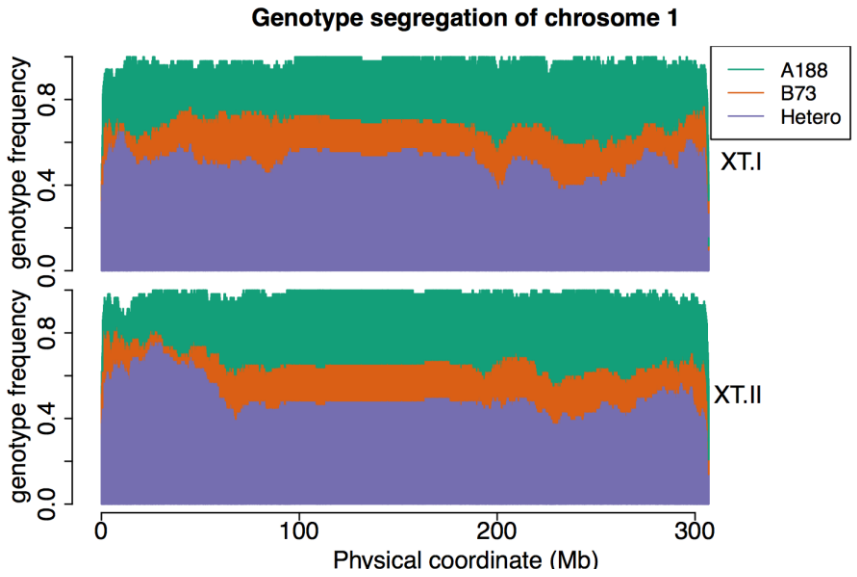

Figure C.1 Genotype segregation of F2 calli on chromosome 1. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white.

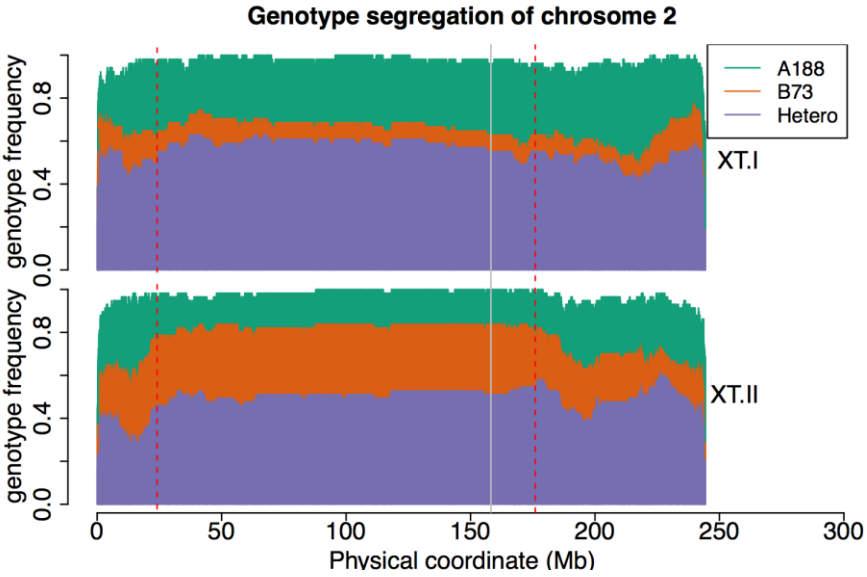

Figure C.2 Genotype segregation of F2 calli on chromosome 1. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white. The red dash line indicated the QTL interval, and the gray vertical line labeled at the QTL position mapped by the GBS seg markers.

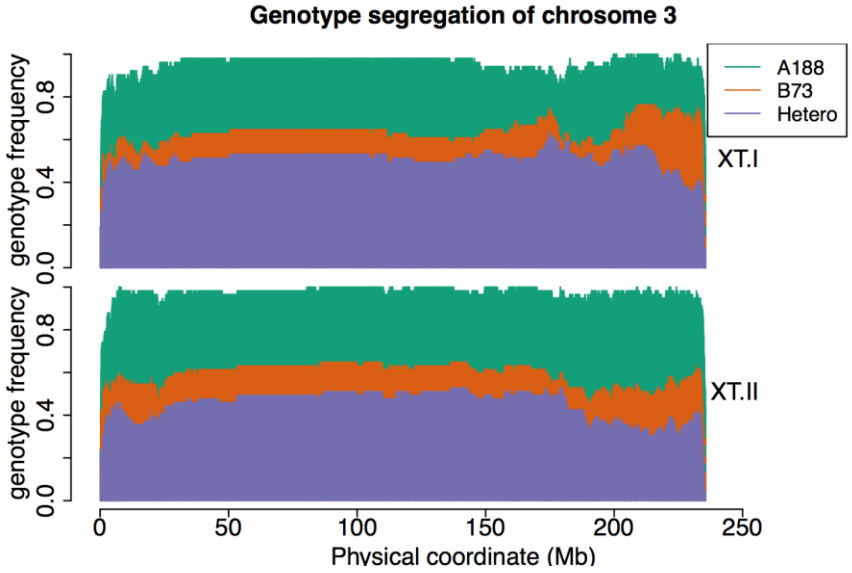

Figure C.3 Genotype segregation of F2 calli on chromosome 3. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white.

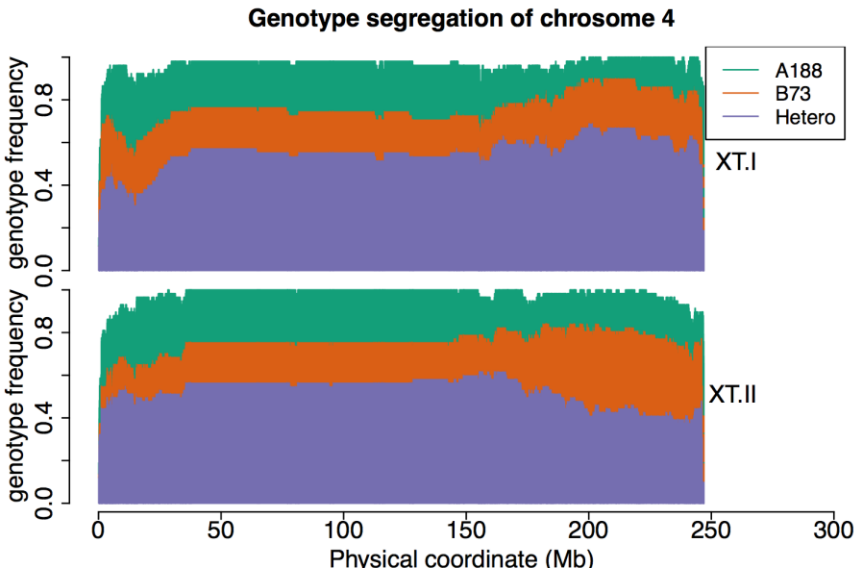

Figure C.4 Genotype segregation of F2 calli on chromosome 4. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white.

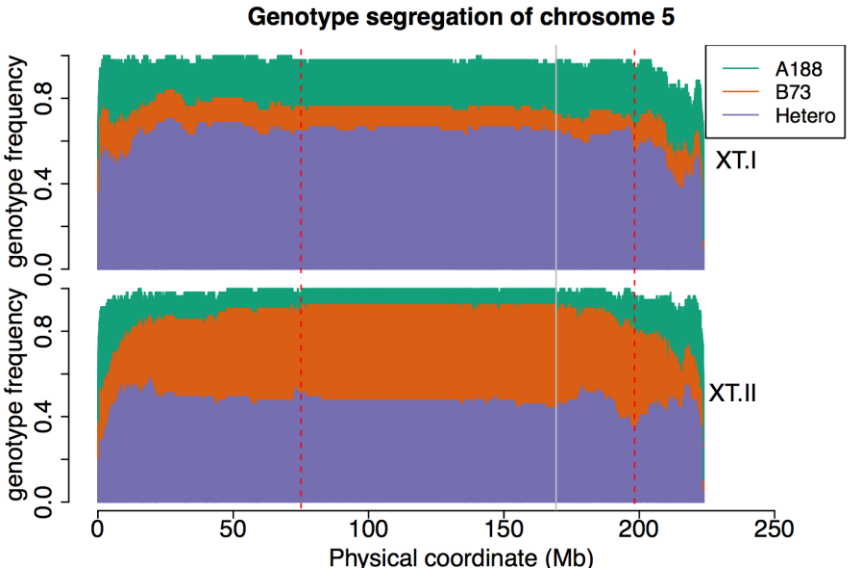

Figure C.5 Genotype segregation of F2 calli on chromosome 5. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white. The red dash line indicated the QTL interval, and the gray vertical line labeled at the QTL position mapped by the GBS seg markers.

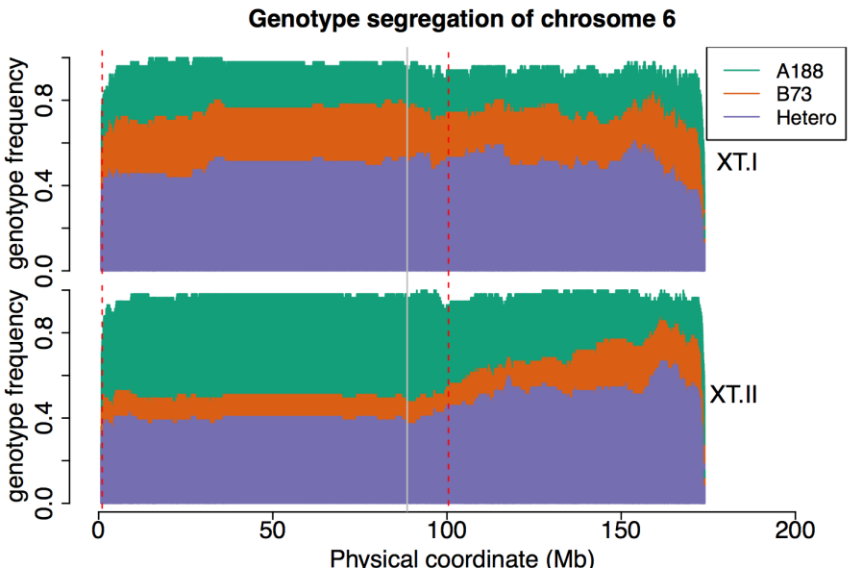

Figure C.6 Genotype segregation of F2 calli on chromosome 6. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white. The red dash line indicated the QTL interval, and the gray vertical line labeled at the QTL position mapped by the GBS seg markers.

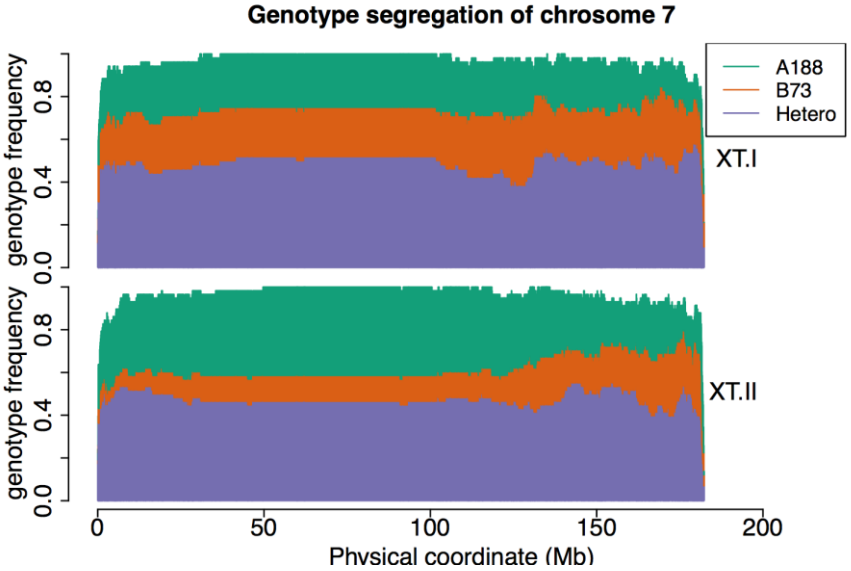

Figure C.7 Genotype segregation of F2 calli on chromosome 7. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white.

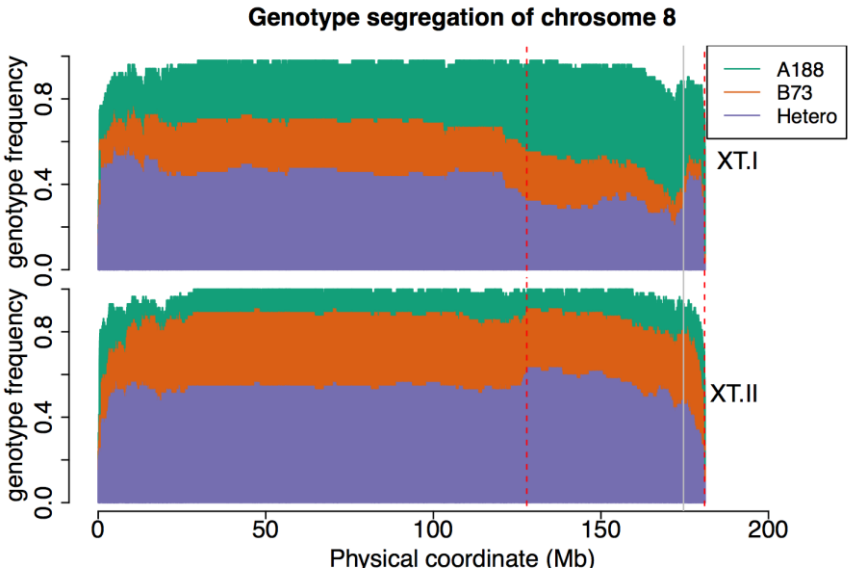

Figure C.8 Genotype segregation of F2 calli on chromosome 8. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white. The red dash line indicated the QTL interval, and the gray vertical line labeled at the QTL position mapped by the GBS seg markers.

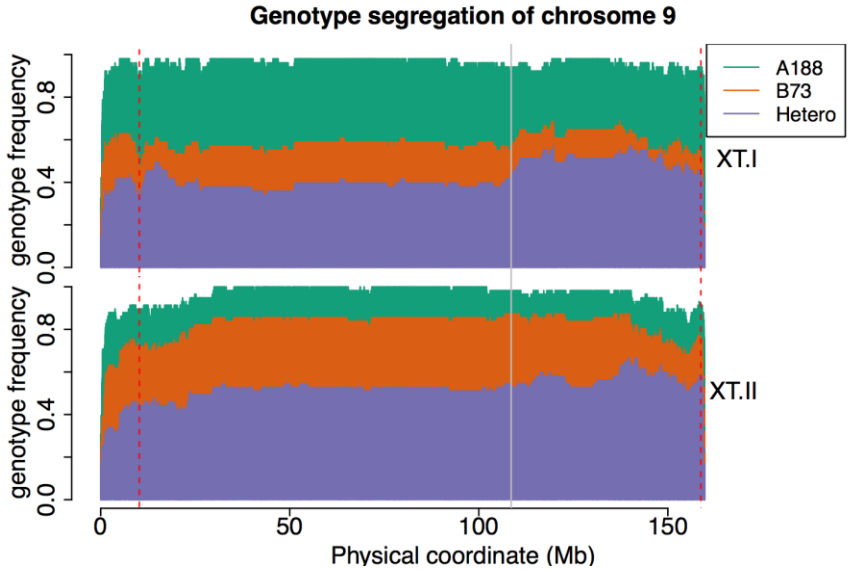

Figure C.9 Genotype segregation of F2 calli on chromosome 9. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white. The red dash line

indicated the QTL interval, and the gray vertical line labeled at the QTL position mapped by the GBS seg markers.

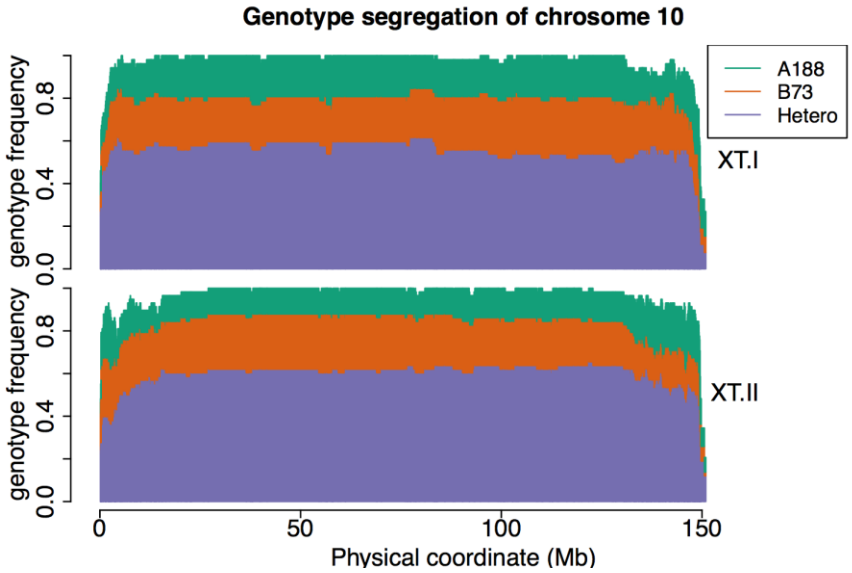

Figure C.10 Genotype segregation of F2 calli on chromosome 10. The upper panel showed the genotype frequency of 60 XT-I individuals, and the bottom panel showed the genotype frequency of 58 XT-II individuals. Each panel resulted from the aggregate of hundreds of vertical lines, and a vertical line represented the genotype distribution of a seg marker. The genotype frequency of heterozygotes, homozygous B73, and homozygous A188 of the seg marker were labelled in purple, orange, and green, and the missing rate of seg marker indicated with white.

Supplemental Tables

Table C.1 The recombination site of Hi-II A and B

Variety	Chromosome	Breakpoint start	Breakpoint end
Hi-II A	1	29788826	31284852
Hi-II A	1	85619028	87989974
Hi-II A	1	180574929	182294746
Hi-II A	2	2806972	4073950
Hi-II A	2	218688090	219601429
Hi-II A	4	195401513	197040039
Hi-II A	4	236305116	237642838
Hi-II A	5	11830286	11934522
Hi-II A	5	48444693	50340857
Hi-II A	5	59281022	63499717
Hi-II A	6	165411893	169000236
Hi-II A	7	2853784	4613286
Hi-II A	7	20297583	20462620
Hi-II A	7	26177306	26703924
Hi-II A	7	137449054	137952706
Hi-II A	9	7721967	8363362
Hi-II A	9	10778140	11243043
Hi-II A	9	13103712	27753901
Hi-II A	9	66812179	68989174
Hi-II A	9	101592249	103077214
Hi-II A	9	108428834	109335212
Hi-II A	10	131155582	133899951
Hi-II B	1	15588668	15967094
Hi-II B	1	30547334	30915591

Hi-II B	1	77152697	77668738
Hi-II B	1	196037123	197952644
Hi-II B	1	269260399	269450192
Hi-II B	1	278639483	282259154
Hi-II B	1	284082304	284325223
Hi-II B	1	285703108	286338963
Hi-II B	2	219340501	219605738
Hi-II B	2	235846855	238068108
Hi-II B	3	220853291	220969622
Hi-II B	3	224973587	225742788
Hi-II B	3	228922713	229396381
Hi-II B	4	7351231	8136195
Hi-II B	4	15020272	15982239
Hi-II B	4	31376411	33181999
Hi-II B	4	36810597	36972978
Hi-II B	4	43673512	44285903
Hi-II B	4	55492642	56277649
Hi-II B	4	154925784	157947748
Hi-II B	4	185175375	185705739
Hi-II B	4	203460228	203811329
Hi-II B	4	205237817	205647602
Hi-II B	4	232560730	233399493
Hi-II B	4	234468486	236643741
Hi-II B	5	4613746	4688978
Hi-II B	5	7506467	8524135
Hi-II B	5	12589478	13379431
Hi-II B	5	61716329	63585220
Hi-II B	5	190396392	191578686

Hi-II B	5	219604659	220998423
Hi-II B	6	108324284	111301573
Hi-II B	7	4964090	5603840
Hi-II B	7	134262506	135482267
Hi-II B	7	147572276	149235763
Hi-II B	7	151981548	152606466
Hi-II B	7	165738518	166119042
Hi-II B	7	168261566	168497523
Hi-II B	8	161887370	162640397
Hi-II B	8	169610649	169917671
Hi-II B	8	176144076	176849751
Hi-II B	9	8120960	9505572
Hi-II B	9	67874714	69278941
Hi-II B	9	91833221	92245606
Hi-II B	9	100840645	102045282
Hi-II B	9	154666389	155603098
Hi-II B	10	15336217	15544393
Hi-II B	10	17707977	18142986
Hi-II B	10	19216382	20024663
Hi-II B	10	105518971	106765463
Hi-II B	10	108135667	109905042
Hi-II B	10	118196289	118820313
Hi-II B	10	125533041	126025041
Hi-II B	10	131247507	133446876

Appendix D - Supplemental data of Chapter 4

Supplemental Figures

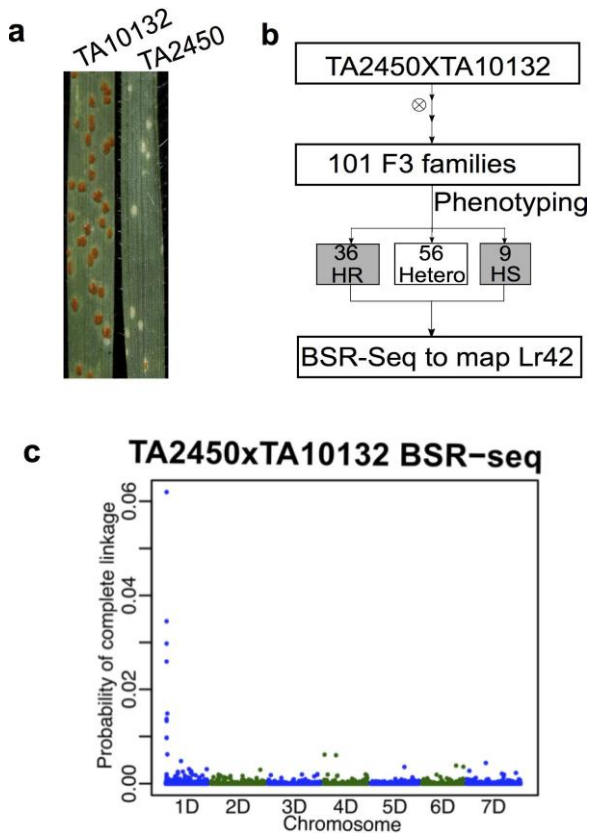

Figure D.1 Genetic mapping of the *Lr42* gene with population TA2450 x TA10132. (**a**) Phenotype of *Ae. tauschii* accessions TA10132 (*Lr42*-) was susceptible (Infection Type = 33+) and TA2450 (*Lr42*+) was hypersensitive flecks (Infection Type = ; to ;1-) at the seedling stage upon inoculation with race PNMRJ. (**b,c**) Genetic mapping of the *Lr42* gene via BSR-seq. Bi-parental population of TA2450 x TA10132 with 101 $F_{2:3}$ families (15 individuals for each family) were phenotyped, and segregated with 36 homozygous resistant (HR) families, 9 homozygous susceptible (HS) families, and 56 heterozygous families. The reason for segregation distortion at the *Lr42* locus in this population is unknown. Among them 26 HR and 9 HS were selected for BSR-seq sequencing.

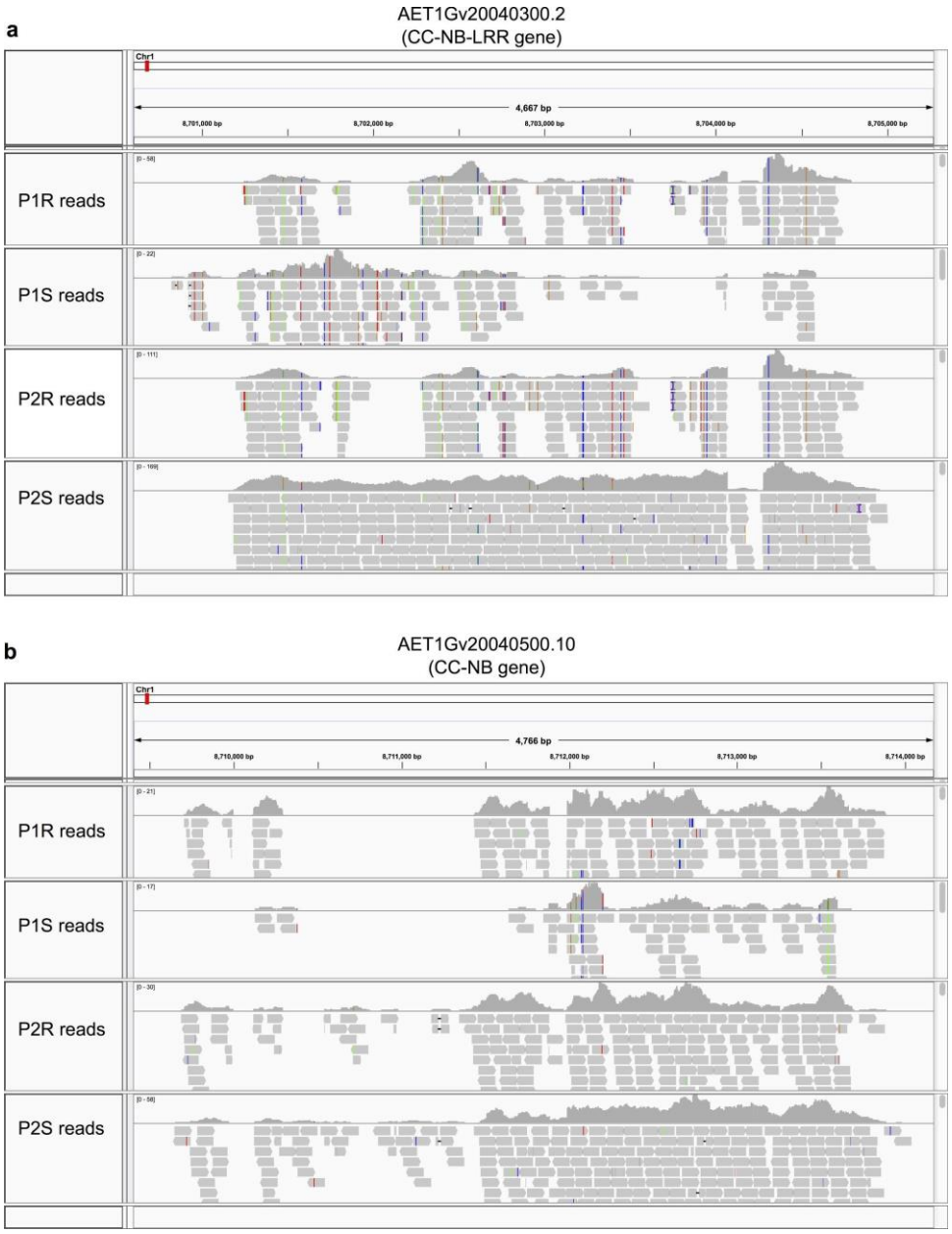

Figure D.2 RNA-Seq read alignments of the two NLR genes in the mapping interval. Integrative Genomics Viewer (IGV) was used to display alignments. In each panel, the top curve is the distribution of read depth along the gene. Non-gray vertical lines represent proportions of alleles at the position that contains multiple types of sequences (alleles). Alignments of reads (horizontal bars) are shown under the curve. P1R and P1S represent RNA-Seq data sets of the resistant pool and the susceptible pool from the mapping population 1, respectively. P2R and P2S represent RNA-Seq data sets of the resistant pool and the susceptible pool from the mapping population 2, respectively. Colored lines in reads highlight polymorphisms between reads and the reference genome.

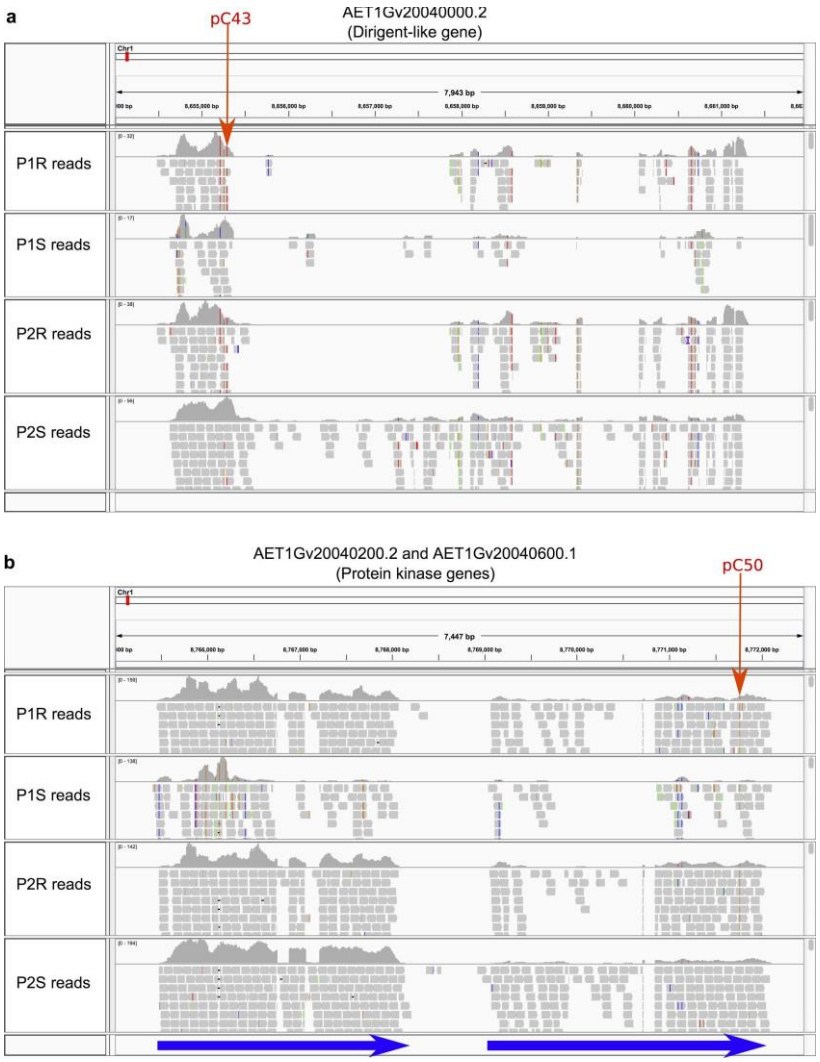

Figure D.3 RNA-Seq read alignments of genes with flanking markers. Integrative Genomics Integrative Genomics Viewer (IGV) was used to display alignments. In each panel, the top curve is the distribution of read depth along the gene. Non-gray vertical lines represent proportions of alleles at the position that contains multiple types of sequences (alleles). Alignments of reads (horizontal bars) are shown under the curve. P1R and P1S represent RNA-Seq data sets of the resistant pool and the susceptible pool from the mapping population 1, respectively. P2R and P2S represent RNA-Seq data sets of the resistant pool and the susceptible pool from the mapping population 2, respectively. Colored lines in reads highlight polymorphisms between reads and the reference genome. (**a**) Read alignment of the gene AET1Gv20040000.2. (**b**) Read alignment of genes *AET1Gv20040200.2* (the first blue arrow) and *AET1Gv20040600.1* (the second blue arrow). The KASP markers pC43 and pC50 were indicated with red arrows.

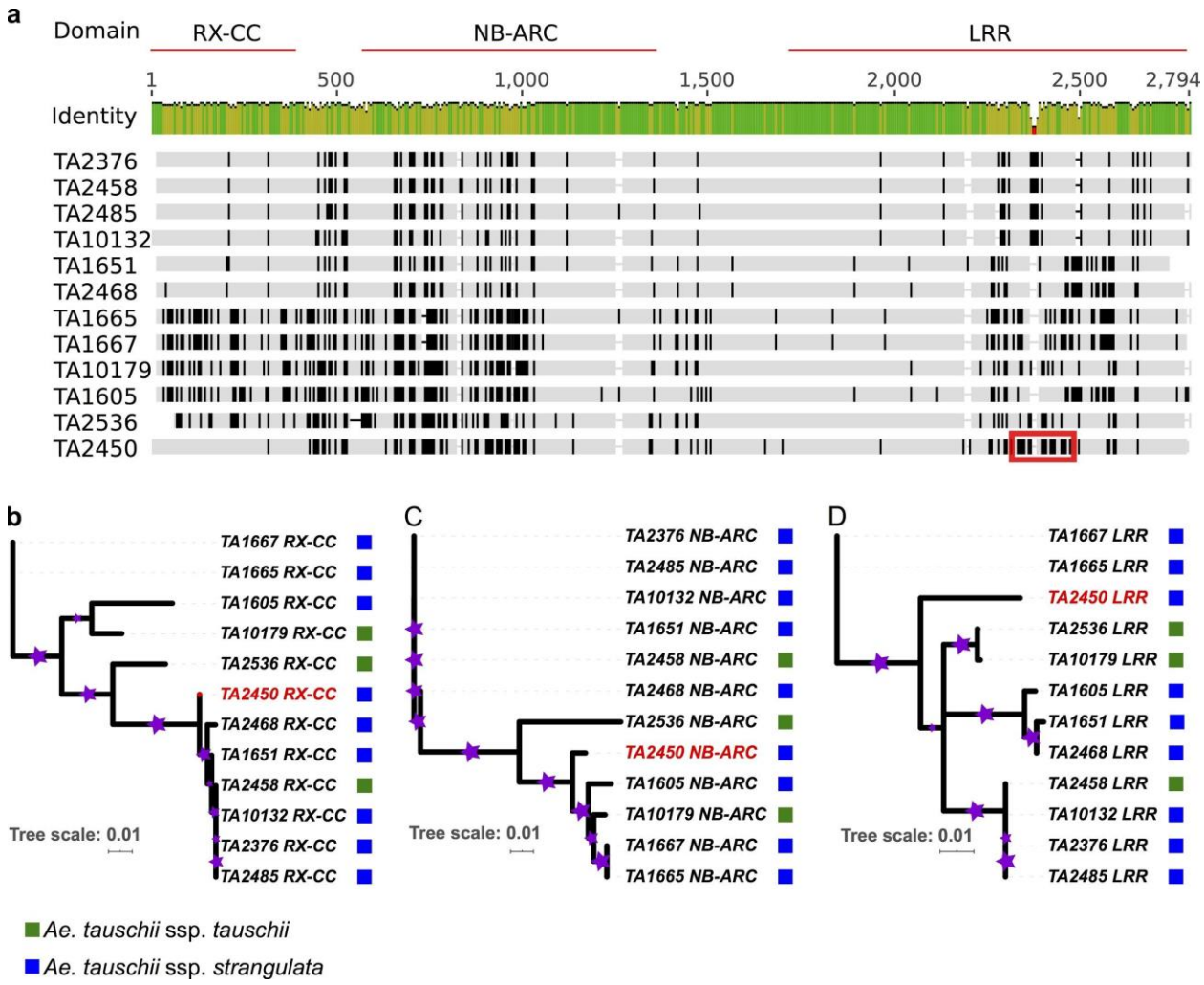

Figure D.4 Sequence and phylogenetic analysis of *Lr42* alleles. (**a**) Multiple alignments using coding sequences of the 12 *Lr42* alleles were generated with Geneious. Three NLR domains are indicated. The identity track has the colors green, yellow and red, representing identities from high to low. Vertical black lines highlight polymorphisms with the consensus sequence. The red box highlights the unique region (*Lr42R*-unique-segment) carried by the *Lr42R* allele. (**b-d**) Separate phylogenetic trees using sequences of the RX-CC, NB-ARC and LRR domains. Green and blue squares at tips represent ssp. *strangulata* (L2) and *tauschii* (L1), respectively. The bootstrap values of clades represented by purple stars were size-coded. The range of sizes from small to large corresponds to the bootstrap values from 54% to 100%.

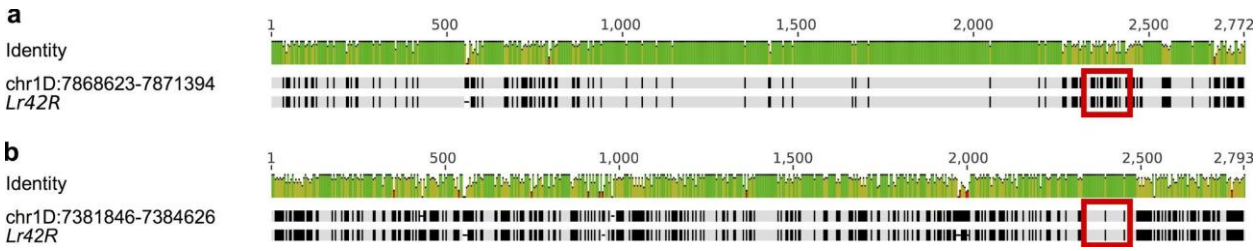

Figure D.5 Pairwise comparison between *Lr42* and homologs in bread wheat. Pairwise alignment was performed using Geneious software. The identity track has the colors green, yellow and red, representing the identities from high to low. Vertical black lines represent polymorphisms between a sequence and the consensus sequence. **(a)** Alignment of the *Lr42R* allele with the homolog with the highest similarity (the top homolog) in the reference genome of CS, which is located at chromosome 1D:7868623-7871394. The identity between *Lr42R* and the top homolog was 93.9%, both of which are considered to be allelic. **(b)** The alignment of the *Lr42R* allele with the CS homolog (1D:7381846-7384626) that carries a highly identical sequence with the *Lr42R*-unique-segment (135/137 bp match from 2,308 to 2,444 bp at the *Lr42R* allele, highlighted with red rectangles). The identity between *Lr42* and this non-allelic homolog was 83.8%.

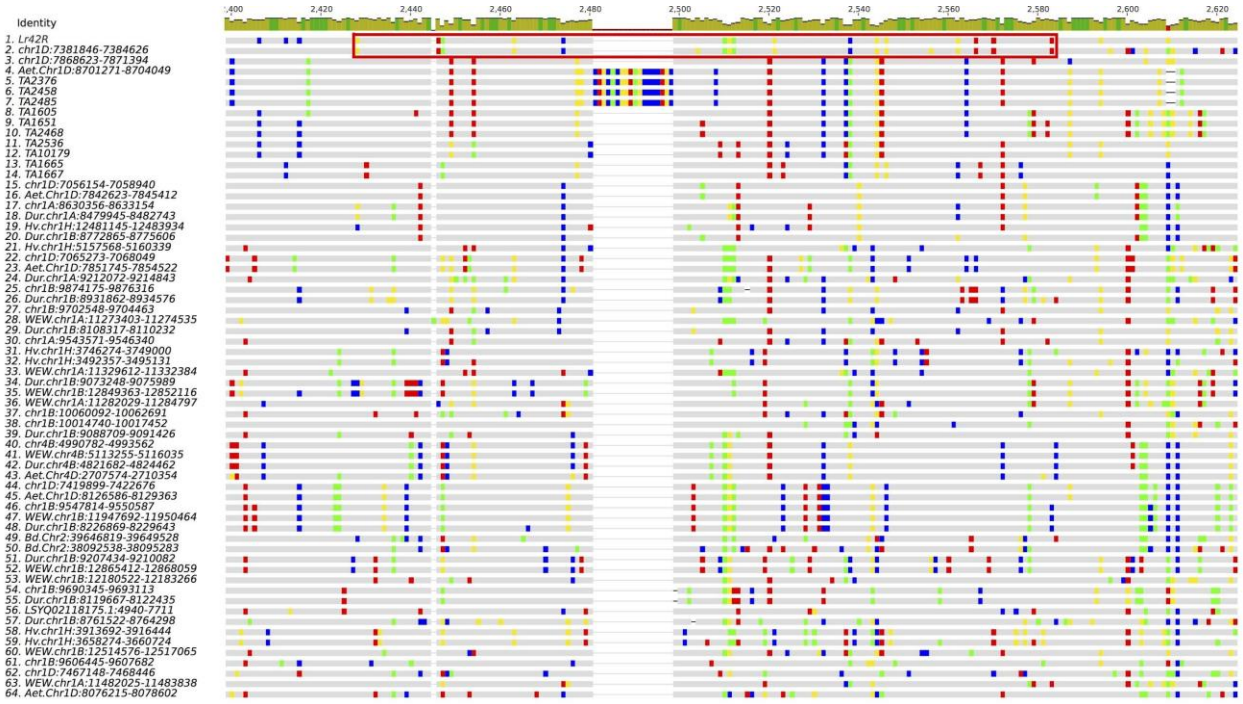

Figure D.6 Multiple alignments of *Lr42* homologs in an LRR region. The *Lr42R* allele was aligned to reference genomes of multiple wheat closely related species using "blastn", and homologs with at least 1 kb matches were extracted. Multiple alignments of sequences of all these homologs and *Lr42* alleles from *Ae. tauschii* were performed using Geneious. The sequences did not contain a region matching the *Lr42R*-unique-segment (highlighted with red rectangle) were manually removed. Green, yellow and red colors of the "identity" track represent the identities of high, medium, and low, respectively. Colors on sequences highlight polymorphisms between a sequence and the consensus sequence. The Sequence IDs beginning with "Aet" are from the reference genome of *Ae. tauschii* (Aet v4.0); sequence IDs beginning with "chr" are from the reference genome *T. aestivum* cv. CS (iwgsc_refseqv1.0); sequence IDs beginning with "Dur" are from the reference genome of *T. turgidum* subsp. *durum* (GCA_900231445.1 Svevo.v1); sequence IDs beginning with "WEW" are from the reference genome of *T. dicoccoides* wild emmer (GCA_002162155.2 WEW v2.0); sequence IDs beginning with "Hv" are from the reference genome of Barley (GCA_901482405.1_ Morex_v1.0,); sequence IDs beginning with "Bd" are from the reference genome of *Brachypodium* (GCF_000005505.3 Brachypodium distachyon v3.0); and sequence IDs beginning with "TA" are *Lr42* alleles from the *Ae. tauschii* accessions.

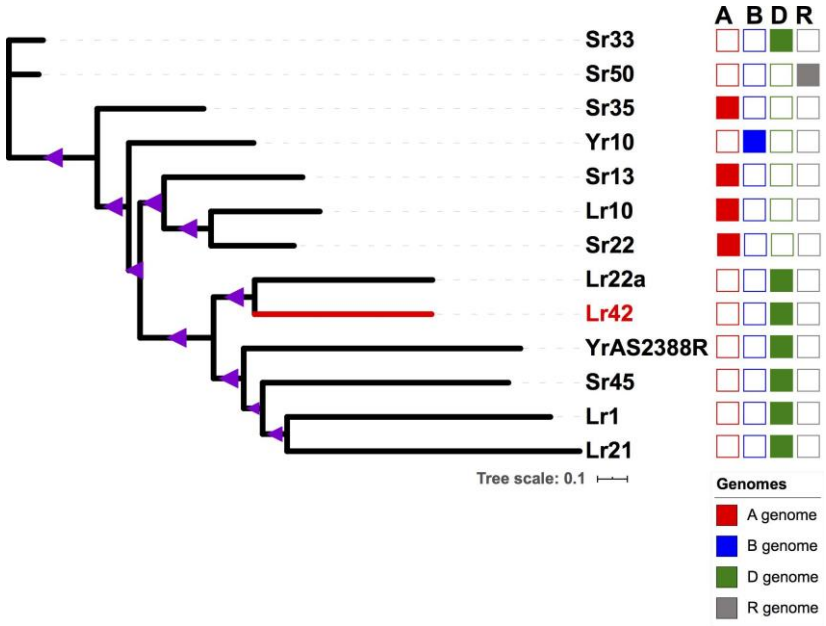

Figure D.7 Phylogenetic tree of cloned wheat rust resistance NLR proteins. Multiple alignments were performed using ClustalW implemented in the Geneious software. The bootstrap values from 56% to 100% are indicated by purple triangles from small to large sizes. Lr22a has the highest identity with Lr42 (31.6%). The genes and their chromosomal locations are: *Lr1*(5DL), *Lr10*(1AS), *Lr21*(1DS), *Lr22a*(2D), *Lr42*(1DS), *Sr13* (R1 haplotype)(6A), *Sr22*(7A), *Sr33*(1DS), *Sr35*(3AL), *Sr45*(1D), *Sr50*(1RS), *YrAS2388R*(4DS), and *Yr10*(1BS).

Supplemental Tables

Table D.1 List of Ae. tauschii accessions used in the study

No.	Accession	Genus	Species	Lineage*	Expected band using the primers LR42_H1F/LR42_H1R	Note	Leaf Rust Infection Score**
1	TA10132	*Aegilops*	*tauschii*	L2	Yes	Susceptible Parent (AL8/78 for reference genome)	3+
2	TA2450	*Aegilops*	*tauschii*	L2	Yes	*Lr42* Resistant parent	;
3	TA2433	*Aegilops*	*tauschii*	L1	No	Susceptible Parent	3+

4	TA1605	*Aegilops*	*tauschii*	L2	Yes	Minicore	2+
5	TA1651	*Aegilops*	*tauschii*	L2	Yes	Minicore	1
6	TA1665	*Aegilops*	*tauschii*	L2	Yes	Minicore	3+
7	TA1667	*Aegilops*	*tauschii*	L2	Yes	Minicore	2
8	TA2376	*Aegilops*	*tauschii*	L2	Yes	Minicore	4
9	TA2458	*Aegilops*	*tauschii*	L1	Yes	Minicore	2
10	TA2468	*Aegilops*	*tauschii*	L2	Yes	Minicore	2+
11	TA2474	*Aegilops*	*tauschii*	L2	Yes***	Minicore	4
12	TA2485	*Aegilops*	*tauschii*	L2	Yes	Minicore	3+
13	TA2536	*Aegilops*	*tauschii*	L1	Yes	Minicore	3
14	TA10179	*Aegilops*	*tauschii*	L1	Yes	Minicore	4
15	TA1578	*Aegilops*	*tauschii*	L1	No	Minicore	3
16	TA1596	*Aegilops*	*tauschii*	L1	No	Minicore	3+
17	TA1631	*Aegilops*	*tauschii*	L1	No	Minicore	3
18	TA1666	*Aegilops*	*tauschii*	L2	No	Minicore	4
19	TA1694	*Aegilops*	*tauschii*	L1	No	Minicore	3
20	TA2374	*Aegilops*	*tauschii*	L1	No	Minicore	4
21	TA2378	*Aegilops*	*tauschii*	L2	No	Minicore	2
22	TA2395	*Aegilops*	*tauschii*	L1	No	Minicore	3
23	TA2413	*Aegilops*	*tauschii*	L1	No	Minicore	3+
24	TA2431	*Aegilops*	*tauschii*	L1	No	Minicore	3
25	TA2448	*Aegilops*	*tauschii*	L1	No	Minicore	2+
26	TA2488	*Aegilops*	*tauschii*	L1	No	Minicore	3
27	TA2508	*Aegilops*	*tauschii*	L1	No	Minicore	4
28	TA2514	*Aegilops*	*tauschii*	L1	No	Minicore	4
29	TA2545	*Aegilops*	*tauschii*	L1	No	Minicore	3
30	TA10099	*Aegilops*	*tauschii*	L1	No	Minicore	4
31	TA10106	*Aegilops*	*tauschii*	L1	No	Minicore	3
32	TA10108	*Aegilops*	*tauschii*	L1	No	Minicore	3+
33	TA10124	*Aegilops*	*tauschii*	L2	No	Minicore	3
34	TA10141	*Aegilops*	*tauschii*	L1	No	Minicore	4
35	TA10144	*Aegilops*	*tauschii*	L1	No	Minicore	3
36	TA10162	*Aegilops*	*tauschii*	L1	No	Minicore	3
37	TA10210	*Aegilops*	*tauschii*	L1	No	Minicore	2+

38	TA10212	*Aegilops*	*tauschii*	L1	No	Minicore	2+

* Lineage data from Singh et al. 2019 except for TA2468 whose lineage information was from the WGRC database.
** At the seedling stage, parental lines TA2450, TA2433, and TA10132 were inoculated with race PNMRJ, which is virulent on *Lr39*. Other *Ae. tauschi* lines were inoculated with a composite of races. Stakman scale scores; ,1, 2, and 2+ are for resistant phenotype, and scores 3, 3+ and 4 are for susceptible phenotype.
*** The expected band was observed in PCR with TA2474 DNA, but the sequence quality was low. The *Lr42* allele was not assembled for accession TA2474.

Table D.2 Avirulence/Virulence specificity of races of *Puccinia triticina* used in study

Race	Avirulent	Virulent
PNMRJ	*2a, 11, 14a, 16, 21, 24, 30, 42*	*1, 2c,3, 3ka, 9, 10, 17, 18, 26, 28, 39, B*
TFBJG	*3ka, 9, 11, 16, 17, 18, 21, 30, 39, 42*	*1, 2a, 2c, 3, 10, 14a, 24, 26, 28*
MBJ/SP	*2a, 2b, 2c, 3ka, 9, 16, 18, 19, 21, 24, 25, (26), 28, 29, 30, 32, 33, 36, 42*	*1, 3, 3bg, 10, 11, 13, 15, 17, 20, 23, 27+31*

Table D.3 KASP markers used for fine mapping in *Ae. tauschii*

Assay	1D_position ***	Orientation	Allele1	Allele2	PCR Size	Primer for Allele1*	Primer for Allele2**	Common primer
pB02	6829289	+	C	G	136	GAGGAGGAGCTTCCCGCCC	GAGGAGGAGCTTCCCGCCG	CCTGGAGCCCTTCGCGAC
pB07	8169715	+	T	C	94	GGCTGGCTGTGTATGTATCTAT	GGCTGGCTGTGTATGTATCTAC	AGCTACGACGACGATGCTG
pB06	8174002	+	C	T	71	TGTGTGGATGATTGGCGGC	TGTGTGGATGATTGGCGGT	CAACCACCATGGCCCAGAT
p12A10	8655251	-	A	G	105	CCTGGAAACGTACGTATGCAT	CCTGGAAACGTACGTATGCAC	GAAAGGCCTCCAGCTCCAG
pB05	8655251	-	A	G	105	CCTGGAAACGTACGTATGCAT	CCTGGAAACGTACGTATGCAC	GAAAGGCCTCCAGCTCCAG
pC43	8655291	+	C	T	105	CCTGGAAACGTACGTATGCAC	CCTGGAAACGTACGTATGCAT	GAAAGGCCTCCAGCTCCAG
pC44	8701579	+	C	T	127	TCCAAGCACGGAGGGTTC	TCCAAGCACGGAGGGTTT	CTCTCTTCCTCCGGTTGGC
p2A04	8704350	+	G	A	60	TGAAGTCCACGAGAAGATGACG	TGAAGTCCACGAGAAGATGACA	TTATCAGACTCTGCGCCGC
pC46	8704527	-	C	G	149	ATTTGCCTTCGAGAACCCG	ATTTGCCTTCGAGAACCCC	GGACGTCGAAGTGTTGCAC
pC30	8712079	-	T	C	96	ACCCACTCAGCAGATCTCA	ACCCACTCAGCAGATCTCG	GCCCAACAAACCATCCACC
pC31	8766128	+	A	G	53	ATGGCGCTCAAAGGCACA	ATGGCGCTCAAAGGCACG	ACTCGCCAAACTGTACCCG
pC50	8771761	-	A	G	106	CACAATTGCCAGCTTCCTCATT	CACAATTGCCAGCTTCCTCATC	CCTCCCAGAGTGGATCTACGA
pC51	8793936	-	T	A	136	CTACAATCCTTCACTG	CTACAATCCTTCACTG	CACCTCGTCGACGGA

						TCTTCA	TCTTCT	AAGG
pC38	8811028	+	A	G	123	TGTGTGGATGTGATCGTTACCA	TGTGTGGATGTGATCGTTACCG	TGCAGCTACTTCAAGGAGCA
pC47	8814844	+	G	A	75	AGATCTCAGACTTTGGCCTG	AGATCTCAGACTTTGGCCTA	TGCCTCTTGCCGCTGTTAA
pC48	8817482	+	T	C	54	TCCGCGCAAGAGCCACCT	TCCGCGCAAGAGCCACCC	TTCTCCATCGCACTACGGC
pB12	8830775	+	A	G	122	TGTAGGCGTTAGATGGAGGGA	TGTAGGCGTTAGATGGAGGGG	CAATGGTAGGTCCGCCACA
pC49	8830775	-	G	A	63	AAACCAAGGGCACAGGGC	AAACCAAGGGCACAGGGT	GCCATCCTGCTGGAAACCT
pC41	8835940	+	G	A	59	AGTAGATACAGGTAGCTTTGCG	AGTAGATACAGGTAGCTTTGCA	CCACGCCTGCCATCTGTAA
pC21	8847059	-	T	C	50	TCTGATGGTGCAGATGTTGAA	TCTGATGGTGCAGATGTTGAG	TCCTTCACACCAACTGGGC
p1A03	8847089	+	C	T	114	GCACCATCAGAAATCCCAACC	GCACCATCAGAAATCCCAACT	TGATGACAGCCAGGCAGTC
pB14	8847089	+	C	T	114	GCACCATCAGAAATCCCAACC	GCACCATCAGAAATCCCAACT	TGATGACAGCCAGGCAGTC
pC22	8847206	+	A	C	115	ACAGCAGAAGCAGAAGCATCA	ACAGCAGAAGCAGAAGCATCC	ACGATGAAGACGCCGTTGT
pC23	8847215	+	G	A	106	GCAGAAGCATCACCAGAAGAG	GCAGAAGCATCACCAGAAGAA	ACGATGAAGACGCCGTTGT
pC24	9887655	+	T	A	127	TGCCATGCTTGGATCAAGGT	TGCCATGCTTGGATCAAGGA	CCCTCAGTACTTGGGCATCC
pC25	9894404	+	G	A	106	AGGAAGGTACGTTATTAGGCCG	AGGAAGGTACGTTATTAGGCCA	GGTGGAAAGAGAACCGCGA
p1A06	9894954	-	C	T	51	TGAAGCTGAAGGACCTCGG	TGAAGCTGAAGGACCTCGA	ACCGAACCCTCTTCAGCTC

pB15	10187788	+	C	G	70	GTGAGGCGGTAGAGG AAGAC	GTGAGGCGGTAGAGG AAGAG	CGCTCACCTTCTTCG ACGT
p1A05	11231012	+	C	G	95	CCGGCCGAATCGAGC CTC	CCGGCCGAATCGAGC CTG	TCCACGAGCTCAAAC CGAG
p1A04	11677167	+	G	A	118	ATAACAGGTGGCCCTC ACG	ATAACAGGTGGCCCTC ACA	ACCGACATGTTCTGC GTCT
p1A02	12012800	+	T	C	133	GCCTCCCATCCCAGAA ACT	GCCTCCCATCCCAGAA ACC	GGGAGGCTGCTTTCC AGAA

* Oligo labeled with fluorophore FAM in KASP master mix (LGC, UK) was needed to add at 5'-ends of primers
** Oligo labeled with fluorophore HEX in KASP master mix (LGC, UK) was needed to add at 5'-ends of primers
***Chromosome 1D position

Table D.4 List of primers and their sequences

No.	Primer name	Sequence (5'-3')	Purpose
1	AET300.2_CDS-F	ATGGCGGAGGCTGTTGTCGGGCAG	For amplification of coding sequences
2	AET300.2_CDS-R	CTAACTTGGCTGAGCTGTGCTGCGC	
3	AET300.2_CDS-BamHIF	GTCGGATCCATGGCGGAGGCTGTTGTCGGGCAG	For plasmid construction to generate transgenic line
4	AET300.2_CDS-BamHIR	ATCGGATCCCTAACTTGGCTGAGCTGTGCTGCGC	
5	AET300.2-Seq500-F	CACTTCACCAGGGATGAGG	For sequencing of coding sequences
6	AET300.2-Seq600-R	CCACACCAGGGTGACTTTG	
7	AET300.2-Seq1000-F	ACAAGAGCATCACTCATGG	
8	AET300.2-Seq1500-F	GGATGCACGATGTCATACG	
9	AET300.2-Seq2000-F	AAGGTGCCTAAGGGCATGC	
10	LR42_H1F	CATGGCGGAGGCTGTTGTC	Alleles/Paralogs amplification
11	LR42_H1R	GTATATCACCTGATCCGTTCACAT	
12	42seq-546R	CAGCCACCGTATCAACCTCT	Allele sequencing
13	42seq-606F	GCCTGGTGTTGGTAAAAC	
14	42seq-1204F	GATGTGATCCCTGATGCTC	
15	42seq-2244R	GGGTCTGAGGTAAATGTAG	
16	42seq-2226F	CTACATTTACCTCAGACCC	
17	actin_F1	CATCCACGAGACGACCTACA	RT-PCR
18	actin_R1	GATCTTCATGCTGCTTGGGG	
19	Lr42-qRT_F5	GGGCTACATTTACCTCAG	
20	Lr42-qRT_R5	CCGTAACACACACAGACA	
21	lr42_1F	TCAGCTCGGATGCCCCAG	
22	lr42_1R	TGCGCCCCATAACTCGATG	
23	Ubi-F	TGGCATATGCAGCAGCTATAT	Transgenic gene screening
24	Seq2R	CCACACCAGGGTGACTTTG	

Table D.5 List of *Lr42*-specific GBS tags

Order	Sequence (5'-3')
1	TGCAGTCACATGCATGTGCACCCAGTGGCAGTCGGTTCTCTCCCTTAATCTGCTCCCTCTAAAC
2	TGCAGCTATCCTTAGCATGTACGTATGTTTCATCTCAAAAAGAAAAAAAGAGCTATCCTTAGCA
3	TGCAGTTGTTTTTTTGAGATGGAGCGCTATTAGATCTAATCCATTAATTACGTTTTTTAAGTTA
4	TGCAGTGATCCGATCCAGGTGTTTGAGGAAGCAGCGCGAGTGTTCAATGATGACTACTTCCTCA
5	TGCAGCTGGGTTTCGATTTTTTGGAGCTTGGCTGCCAGGCGGCGCCAGGTCTTGATATGCTTGA
6	TGCAGATGTTCTGCAAATTATAAACTACTCACTATACCATGAAAAAATTTAGGGGGGCACCATA
7	TGCAGGACTAGCTAGGTACGAAGAAGCAATTGTTTACCTGGTGCCGCCAGAGGGCAGGATGATT
8	TGCAGGCAGGTTGACGAGGGCATCTACTGTACACAAGGCCCGCAAATCAAATCAAGTCCAATTC
9	TGCAGTTGAAGGGCACAATTTTTTTGTTGGAGATTATAGTACGATAAATTCAGCTTGTACTCTA
10	TGCAGCACATTGACCTAATATACGAGGAACAAATATAGGAAGCTATTTATGCCTTCCAGATTCA
11	TGCAGGTGGTTGAGTATGCATCCCCAGCACTAACCCCAGCCCCAACCACACATGATACCTGTAG
12	TGCAGCTAGTTAAGTAACATTAGTAGGTACGCCTTTAATTTCTTGTTGAATACTCGCAAGTTTA
13	TGCAGGGTGTGTAATTTGAACAAAACCACGACGAGTAATTTAGAACGGAGGGAGTATATCATAT
14	TGCAGATCGATGGTGAGTGCGCCGACCGAGAGAAGGAAGGAGCTGCGGGGGATGCGGGGACGAG

Table D.6 KASP markers useful for marker-assisted selection in hexaploid wheat

Marker	Susceptible allele*	Resistant allele**	Common_primer	PCR size	Marker position	Note
Lr42_pD1	ACAAGAAAGAGCAGACCA	GCAAGAAAGCGTAGACCC	CCCTTTCTTGGCTTGGTTTAG	190	In gene, LRR region	dominant in most cases
Lr42_pD2	CTGCGTAAAAATCCAAGCGCT	CTCTGCGTAAAAATTCAATCTCC	CCCTTTCTTGGCTTGGTTTAG	251	In gene, LRR region	dominant in most cases
pC43***	CCTGGAAACGTACGTATGCAC	CCTGGAAACGTACGTATGCAT	GAAAGGCCTCCAGCTCCAG	105	46kb to the gene	co-dominant

* Oligo labeled with fluorophore FAM in KASP master mix (LGC, UK) was needed to add at 5'-ends of primers
** Oligo labeled with fluorophore HEX in KASP master mix (LGC, UK) was needed to add at 5'-ends of primers
***This marker was also used in fine mapping; 46kb to *Lr42*

Table D.7 LRR repeats in the *Lr42* resistance allele

Repeat order	Amino acid sequence	Start*	End*	Length (aa)
repeat1	LRALHVFESYIDIGLLKPILTSSSL	551	575	25
repeat2	LSTLDLQGTHIKMLPNEVFDLFN	576	598	23
repeat3	LRYLGLRDTKIESMPAAVGRLQN	599	621	23
repeat4	LQVLDAYDSKLTYLPNSVVKLQK	622	644	23
repeat5	LRYLYAGTSKDSIRGVKVPKGMQH	645	668	24
repeat6	LAGLHALQSVKATPEFLHEAAA	669	690	22
repeat7	LTELRTFDVCNVQSEHSAYLSNAITKMSH	691	719	29
repeat8	LVHLEIDAAAENDVLRLEGLHLPQT	720	744	25
repeat9	LSWLGLAGQLEKTTMPQLFSSWSHLDS	745	771	27
repeat10	LTRLYLAFSSIDEQTFSCLCVLRG	772	795	24
repeat11	LRFLALREAFEGRRLNFYAESFPE	796	819	24
repeat12	LRHLEIWGAAQLSQVRIEKGAMQN	820	843	24
repeat13	LIELFFTDCPDLRFLPDGIEHLAG	844	867	24
repeat14	LEKLGLIDTSEELIEKLRQDRDSDACSKDLMKISHIRM VGVQLGQKGLCERIR	868	920	53

* positions on the Lr42 protein sequence

Table D.8 Yield traits of wheat lines with and without *Lr42R*

Trait	Number of *Lr42+*	Number of *Lr42-*	Effect[%]	p-value
Test weight	105	146	0.3	0.24
Grain yield Severe drought	105	146	0.21	0.282
Thousand Kernel weight	105	146	0.79	0.292
Grain yield Optimum irrigation Bed planting	105	146	-0.09	0.414
Grain yield Drought	105	146	0.09	0.521
Grain yield Optimum irrigation Flat planting	105	146	-0.05	0.614
Grain yield Late-sown heat stress	105	146	-0.04	0.77

[%] mean difference betwen *Lr42+* and *Lr42-* lines

Table D.9 The resutls of KASP assay in CIMMITY breeding lines with Lr42-pD1 and Lr42-pD2

Order	CIMMY_GID	Pedigree	Lr42-pD1%	Lr42-pD %	*Lr42R*
1	6937922	QUAIU #1*2/MUNAL #1	R	R	Present
2	6932911	QUAIU*2/DANPHE	R	R	Present
3	6938530	WHEAR/SOKOLL *2//QUAIU #1	R	R	Present
4	6935970	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/ TRAP//KAUZ/5/BECARD	R	R	Present
5	6935972	QUAIU #1/3/PBW343*2/KUKUNA// PBW343*2/KUKUNA	R	R	Present
6	6931606	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/ TRAP//KAUZ/5/KINGBIRD #1	R	R	Present
7	6936896	QUAIU/FRNCLN	R	R	Present
8	6939176	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ*2/5/PRL/2*PASTOR/4/CHOIX/...	R	R	Present
9	6939666	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ*2/5/PRL/2*PASTOR/4/CHOIX/...	R	R	Present
10	6939180	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/ TRAP//KAUZ*2/5/MUNAL #1	R	R	Present
11	6939492	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ/5/TRCH/SRTU//KACHU	R	R	Present
12	6939303	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ*2/5/BECARD	S	S	Absent
13	6939089	ND643/2*WAXWING/5/BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ	R	R	Present
14	6939127	QUAIU//KIRITATI/2*TRCH	R	R	Present
15	6939174	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ*2/5/PRL/2*PASTOR/4/CHOIX/...	R	R	Present
16	6939181	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ*2/5/MUNAL #1	R	R	Present
17	6939033	BLOUK#1/4/WBLL1/KUKUNA//TACUPETO F2001/3/ UP2338*2/VIVITSI	R	R	Present
18	6939818	QUAIU/3/EMB16/CBRD//CBRD/4/QUAIU #1	R	R	Present

19	6939343	PFAU/MILAN/3/BABAX/LR42//BABAX*2/4/NIINI #1	S	S	Absent
20	6939304	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ*2/5/BECARD	R	R	Present
21	7313634	TACUPETO F2001/SAUAL//BLOUK #1/3/SAUAL/YANAC//SAUAL/4/TACUPETO F2001/SAUAL//...	R	R	Present
22	7396083	QUAIU #1/3/KINGBIRD #1//INQALB 91*2/TUKURU	R	R	Present
23	7396161	SAUAL/YANAC//SAUAL/3/BECARD/QUAIU #1	R	R	Present
24	7310720	BECARD/QUAIU #1//BORL14	R	R	Present
25	7396738	WHEAR/KUKUNA/3/C80.1/3*BATAVIA//2*WBLL1/4/QUAIU/5/BORL14	R	R	Present
26	7311326	QUAIU#1/5/KIRITATI/4/2*BAV92//IRENA/KAUZ/3/HUITES/6/BECARD/QUAIU #1	R	R	Present
27	7398322	WAXWING/4/BL1496/MILAN/3/CROC_1/AE.SQUARROSA(205)//KAUZ/5/FRNCLN/6/...	R	R	Present
28	7401507	TACUPETO F2001/BRAMBLING//KIRITATI/3/FRAN COLIN #1/BLOUK #1/4/FRANCOLIN #1/...	R	R	Present
29	7401659	QUAIU/YANAC//FRANCOLIN #1/BLOUK #1/3/FRANCOLIN #1/BLOUK #1	R	R	Present
30	7305876	FRANCOLIN #1/BLOUK #1/3/ KINGBIRD #1//INQALB 912/TUKURU/4/FRANCOLIN #1/BLOUK #1	R	R	Present
31	7401768	FRANCOLIN #1/BLOUK #1/3/PBW343*2/KUKUNA*2//FRTL/PIFED/4/FRANCOLIN #1/BLOUK #1	R	R	Present
32	7401788	FRANCOLIN #1/BLOUK #1*2/4/KACHU #1//PI 610750/SASIA/3/KACHU	R	R	Present
33	7401811	FRANCOLIN #1/BLOUK #1*2/3/SAUAL/YANAC//SAUAL	?	R	Present
34	7400983	OASIS/SKAUZ//4*BCN/3/PASTOR/4/KAUZ*2/YACO//KAUZ/5/2*QUAIU #3/6/BECARD/QUAIU #1	R	R	Present
35	7400990	WBLL1*2/BRAMBLING//QUAIU/6/BABAX/LR42//BABAX*2/3/KUKUNA/4/CROSBILL #1/5/BECARD	R	R	Present

36	6681676	QUAIU #1/SUP152	R	R	Present
37	6681817	SUP152/QUAIU #2	R	R	Present
38	6568414	QUAIU/3/KIRITATI//PBW65/2*SERI.1B/4 /DANPHE #1	R	R	Present
39	6682188	QUAIU #1/2*SUP152	R	R	Present
40	6682697	QUAIU #2//ND643/2*WBLL1/3/ ND643/2*WBLL1	R	R	Present
41	6680725	QUAIU #1/BECARD	R	R	Present
42	6679874	FRNCLN/QUAIU//FRANCOLIN #1	R	R	Present
43	6681677	QUAIU #1/SUP152	S	?	Absent
44	6684073	QUAIU #2*2/TINKIO #1	R	R	Present
45	6570205	FRANCOLIN #1/BLOUK #1	R	R	Present
46	6684819	BLOUK #1//WBLL1*2/BRAMBLING /3/WBLL1*2/BRAMBLING	R	R	Present
47	6569678	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/ 3/KAUZ*2/TRAP//KAUZ/5/PRL/2*PASTO R/4/CHOIX/STAR/3/...	R	R	Present
48	6684655	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/ 3/KAUZ*2/TRAP//KAUZ/5/WAXWING*2/ KRONSTAD F2004	R	R	Present
49	6570210	FRANCOLIN #1/BLOUK #1	S	S	Absent
50	6569768	VILLA JUAREZ F2009/5/BABAX/LR42//BABAX*2/4/SNI/ TRAP#1/3/KAUZ*2/TRAP//KAUZ	R	R	Present
51	6684630	WHEAR/ KIRITATI/3/C80.1/ 3*BATAVIA//2*WBLL1/5/BABAX/LR42// BABAX*2/4/SNI/TRAP#1/3/...	R	R	Present
52	6684527	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/ 3/KAUZ*2/TRAP//KAUZ/5/PRL/2*PASTO R/4/CHOIX/STAR/3/...	R	R	Present
53	6684652	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/ 3/KAUZ*2/TRAP//KAUZ/5/WAXWING*2/ KRONSTAD F2004	S	S	Absent
54	6684622	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/ 3/KAUZ*2/TRAP//KAUZ/5/SAAR*2//PBW 343*2/KUKUNA	R	R	Present
55	6569847	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/ 3/KAUZ*2/TRAP//KAUZ/5/TACUPETO F2001*2/BRAMBLING	R	R	Present

56	6684394	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ2/TRAP//KAUZ/5/WHEAR/SOKOLL	R	R	Present
57	6569872	PFAU/WEAVER*2//TRANSFER#12,P88.272.2/5/BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/...	R	R	Present
58	6569899	VEE#8//JUP/BJY/3/F3.71/TRM/4/BCN/5/KAUZ/6/MILAN/KAUZ/7/SKAUZ/PARUS//PARUS/8/BABAX/...	R	R	Present
59	6569728	KRL19/5/BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ	R	R	Present
60	6569740	BJY/COC//PRL/BOW/3/FRTL/5/BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ	R	R	Present
61	6690406	BABAX/LR42//BABAX/3/ER2000/4/NAVJ07	S	S	Absent
62	3822784	PRL/2*PASTOR	S	S	Absent
63	6085788	QUAIU #1	R	R	Present
64	6333158	SUP152/BAJ #1	S	S	Absent
65	6176013	BORLAUG100 F2014	S	S	Absent
66	5106304	BAJ #1	S	S	Absent
67	4755014	KACHU #1	S	S	Absent
68	14853	PASTOR	S	S	Absent
69	5390612	SUPER152	S	S	Absent
70	6922197	KACHU/SOLALA	S	S	Absent
71	6977141	ARULA #1	S	S	Absent
72	NA	TA2450	R	R	Present
73	NA	TA10132	S	S	Absent
74	NA	KS93U50	R	R	Present
75	NA	Morocco	S	S	Absent
76	NA	NIL_R1[%%]	R	?	Present
77	NA	NIL_S1[%%]	S	S	Absent
78	NA	NIL_R3[%%]	R	R	Present
79	NA	NIL_S3[%%]	S	S	Absent

[%]Genotype result using the two KASP markers. R=Resistant genotype; S=Susceptible genotype; ?=missing data
[%%] near-isogenic lines (F5) derived from KS93U50 and Morocco; R and S represent resistant and susceptible lines, respectively. NIL-R1 is isogenic to NIL-S1 and NIL-R3 is isogenic to NIL-S3.

Table D.10 MADA motif sequences of known wheat rust resistance NLRs

NLR/motif	Motif sequence																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
MADA	M	A	D	A	x	V	S	F	x	V	x	K	L	x	x	L	L	x	x	E	x
Lr42	M	A	E	A	V	V	G	Q	L	V	V	T	L	G	E	A	L	A	K	E	A
Lr22a	M	A	E	A	A	L	L	L	V	T	T	K	I	G	K	A	V	A	T	E	T
Lr1	M	A	A	A	L	G	S	A	A	T	L	L	G	K	V	F	T	M	L	S	A
Lr21	M	A	T	A	W	D	V	A	S	V	G	W	S	M	V	V	L	G	W	L	V
Lr10	M	A	P	C	L	V	S	A	S	T	G	A	M	G	S	L	L	T	K	L	E
Sr45	M	A	E	F	V	V	R	P	L	V	S	T	L	M	N	T	A	S	S	Y	L
Sr13	M	A	E	F	V	V	R	P	L	V	S	T	L	M	N	T	A	S	S	Y	L
Sr22	M	A	E	V	L	V	S	A	S	T	G	A	M	G	S	L	L	R	K	L	G
YrAS2388R	M	A	G	V	L	D	A	L	A	S	Y	V	T	N	M	L	T	E	M	A	K
Sr33	M	D	I	V	T	G	A	I	A	K	L	I	P	K	L	G	E	L	L	V	G
Sr35	M	E	I	A	M	G	A	I	G	S	L	L	P	K	L	G	E	L	L	I	G
Yr10	M	E	V	V	T	G	A	M	S	T	L	L	P	L	L	G	D	L	L	K	E
Sr50	M	N	I	V	T	G	A	M	G	S	L	I	P	K	L	G	E	L	L	M	D

First 21aa of each NLR protein were listed.

Table D.11 CIMMYT varieties developed from *Lr42* wheat lines

Country	Name	Pedigree
Afghanistan	Wafer 15	BABAX/LR42//BABAX*2/3/TUKURU
Afghanistan	Koshan 09	BABAX/LR42//BABAX*2/3/VIVITSI (=QUAIU)
Ethiopia	Gambo	BABAX/LR42//BABAX*2/3/VIVITSI (=QUAIU)
Iran	Aftab	THELIN/3/BABAX/LR42//BABAX/4/BABAX/LR42//BABAX
Kenya	Kenya Peacock	QUAIU/3/PGO/SERI/BAV92
Mexico	RSM Norman F2008	BABAX/LR42//BABAX
Nigeria	Norman	BABAX/LR42//BABAX
Rwanda	Reberaho	BABAX/LR42//BABAX*2/4/SNI/TRAP#1/3/KAUZ*2/TRAP//KAUZ
Rwanda	Rengerabana	BABAX/LR42//BABAX*2/3/TUKURU
Turkey	Ekinoks	BABAX/LR42//BABAX*2/3/VIVITSI (=QUAIU)
Kenya	Robin	BABAX/LR42//BABAX*2/3/TUKURU