Genetic diversity of wheat wild relative, *Aegilops tauschii*, for wheat improvement

by

Narinder Singh

B.Sc., Punjab Agricultural University, 2010
M.S., New Mexico State University, 2013

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

2017

Abstract

Wheat is perhaps the most important component in human diet introduced since the conception of modern agriculture, which provides about 20% of the daily protein and calorie intake to billions of people. Adaptable to wide range of climates, wheat is grown worldwide, lending it the potential to mitigate the imminent risk of food security for future population of 9.5 billion people.

For developing improved crop varieties in the future, genetic diversity is a key factor in plant breeding. Constraints in wheat evolution and artificial selection practices have resulted in erosion of this ingredient in elite germplasm. However, wheat wild relatives, such as *Ae. tauschii*, D-genome donor of wheat, are a storehouse for unexploited genetic diversity that can be used for improving wheat for disease and insect resistance, yield, quality, and tolerance to abiotic stresses.

More than 1700 genebanks around the world hold over 7 million accessions of these wild relatives. These genebanks are expensive to maintain, therefore, efficient curation is necessary. We developed and implemented a protocol to identify duplicate accessions using genomic tools. Implementing this approach with three genebanks, we identified over 50% duplicated accessions across genebanks. There are over a million *Triticeae* accessions held collectively, and it is likely as more number of genebanks are tested, there will be decreasing number of unique accessions.

Selecting and utilizing the wild genetic diversity is no easy task. Historically, breeders and geneticists have chosen the accessions primarily based on associated phenotypic data. Unless focusing on a targeted trait, this practice is imperfect in capturing the genetic diversity with some other limitations, such as confounding phenotypic data with the testing environment. Utilizing next-generation sequencing methods, we selected a MiniCore consisting of only 40 accessions

out of 574 capturing more than 95% of the allelic diversity. This MiniCore will facilitate the use of genetic diversity present in *Ae. tauschii* for wheat improvement including resistance to leaf rust, stem rust, Hessian fly, and tolerance to abiotic stresses.

Hessian fly is an important insect pest of wheat worldwide. Out of 34 known resistance genes, only six have been mapped on the D sub-genome. With swift HF evolution, we need to rapidly map and deploy the resistance genes. Some of the undefeated HF resistance genes, such as *H26* and *H32*, were introgressed from *Ae. tauschii*. In this study, we mapped three previously known genes, and a new gene from *Ae. tauschii* accession *KU2147*. Genes were mapped on chromosomes 6B, 3D, and 6D. Further, identification and cloning of resistance genes will enhance our understanding about its function and mode of action.

In conclusion, wild wheat relatives are genetically diverse species, and utilizing the novel genetic diversity in *Ae. tauschii* will be fruitful for wheat improvement in the wake of climate change to ensure future food security to expected 2 billion newcomers by 2050.

Genetic diversity of wheat wild relative, *Aegilops tauschii*, for wheat improvement

by

Narinder Singh

B.Sc., Punjab Agricultural University, 2010
M.S., New Mexico State University, 2013

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

2017

Approved by:

Major Professor
Jesse Poland

Copyright

Abstract

Wheat is perhaps the most important component in human diet introduced since the conception of modern agriculture, which provides about 20% of the daily protein and calorie intake to billions of people. Adaptable to wide range of climates, wheat is grown worldwide, lending it the potential to mitigate the imminent risk of food security for future population of 9.5 billion people.

For developing improved crop varieties in the future, genetic diversity is a key factor in plant breeding. Constraints in wheat evolution and artificial selection practices have resulted in erosion of this ingredient in elite germplasm. However, wheat wild relatives, such as *Ae. tauschii*, D-genome donor of wheat, are a storehouse for unexploited genetic diversity that can be used for improving wheat for disease and insect resistance, yield, quality, and tolerance to abiotic stresses.

More than 1700 genebanks around the world hold over 7 million accessions of these wild relatives. These genebanks are expensive to maintain, therefore, efficient curation is necessary. We developed and implemented a protocol to identify duplicate accessions using genomic tools. Implementing this approach with three genebanks, we identified over 50% duplicated accessions across genebanks. There are over a million *Triticeae* accessions held collectively, and it is likely as more number of genebanks are tested, there will be decreasing number of unique accessions.

Selecting and utilizing the wild genetic diversity is no easy task. Historically, breeders and geneticists have chosen the accessions primarily based on associated phenotypic data. Unless focusing on a targeted trait, this practice is imperfect in capturing the genetic diversity with some other limitations, such as confounding phenotypic data with the testing environment. Utilizing next-generation sequencing methods, we selected a MiniCore consisting of only 40 accessions

out of 574 capturing more than 95% of the allelic diversity. This MiniCore will facilitate the use of genetic diversity present in *Ae. tauschii* for wheat improvement including resistance to leaf rust, stem rust, Hessian fly, and tolerance to abiotic stresses.

Hessian fly is an important insect pest of wheat worldwide. Out of 34 known resistance genes, only six have been mapped on the D sub-genome. With swift HF evolution, we need to rapidly map and deploy the resistance genes. Some of the undefeated HF resistance genes, such as *H26* and *H32*, were introgressed from *Ae. tauschii*. In this study, we mapped three previously known genes, and a new gene from *Ae. tauschii* accession *KU2147*. Genes were mapped on chromosomes 6B, 3D, and 6D. Further, identification and cloning of resistance genes will enhance our understanding about its function and mode of action.

In conclusion, wild wheat relatives are genetically diverse species, and utilizing the novel genetic diversity in *Ae. tauschii* will be fruitful for wheat improvement in the wake of climate change to ensure future food security to expected 2 billion newcomers by 2050.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am indebted to Dr. Jesse Poland for his constant guidance and teaching. As a true mentor, and an inspiration, he took me on so that I could learn new things regardless that I didn't know anything about genomics beforehand. It wouldn't be an exaggeration if I say, I couldn't have asked for a better advisor. I would also like to thank Dr. Bikram Gill for imparting in me the old-school wisdom, and teaching me the value of going to the greenhouse and "talking to plants." Further, the intellectual contributions by my other committee members Drs. Allan Fritz and Geoffrey Morris, were instrumental to my overall development. I would also like to extend my gratitude to Drs. Ming-Shun Chen, Sanzhen Liu, Bernd Friebe, and Dal-Hoe Koo, and Anu Aulakh for helping me at times and answering random questions.

I cannot thank enough all the help from our lab and field technicians. A special thanks to Shuangye Wu for accommodating my emergency genotyping requests, John Raupp for fulfilling ad hoc seed requests, and Duane Wilson and Byron Evers for taking care of my greenhouse and field needs. All the help I received from the undergraduate and graduate students in planting, phenotyping, harvesting, and threshing was significantly contributory towards my PhD.

Organization is a skill that can either be learned or done better by someone who is good at it. In that regards, I was fortunate enough to be surrounded by super-organizers Haley Ahlers and Bo Winters, who were only a quick office-stop away to take care of ordering, traveling and just about anything else. I also appreciate all the help from the Plant Pathology office staff over the course of my PhD.

Finally, for the love and moral support, I thank all my friends here and back home, who believed in me, and made it possible for me to strive and push through the hard times to achieve this lifelong goal.

Dedication

I dedicate this *magnum opus* (some will call it just a dissertation) to my family for their continuous love and support, and enabling me to reach where I am today. Being the first generation to enter the college, they will be happy to see me at the zenith of the academic pursuit. A special dedication to my fiancée (soon to be wife) for an unexpected but motivational ultimatum. Turns out that "we won't marry until you graduate" is a powerful motivator that surely helped me wrap up my dissertation in a timely manner.

On a serious note, I would like to dedicate this dissertation to all the researchers around the world who are contributing towards making the world a better place, and especially the farmers who are the unsung heroes of the battle against food security.

"The day will come
when no one dies of hunger
and world is a better place
the day will come
when farmers won't commit suicide
because their crops will never fail
that day, my inner scientist
will rest, and sleep like a baby
until then
I will keep fighting"

Chapter 1 - *Aegilops tauschii* and Wheat Improvement

Origin of wheat

In the fertile crescent near Caspian Sea, concurrent with the rise of agriculture, a rare natural hybridization event between tetraploid wheat (*Triticum turgidum*; 2n = 4x = 28; AABB) and a diploid goat grass (*Aegilops tauschii*; 2n = 2x = 14; DD), gave rise to the allohexaploid known as bread wheat (*Triticum aestivum*; 2n = 6x = 42; AABBDD) (Kihara 1944; McFadden & Sears, 1946; Lev-Yadun *et al.*, 2000; Marcussen *et al.*, 2014). Due to its superior phenotype and bread making qualities over wild relatives, hexaploid wheat was favorably adopted worldwide.

With future population projected to surpass 9.5 billion by 2050, global food security is at risk (United Nations *et al.*, 2015). Developing countries are at higher risk because of the predicted population growth is greatest in these regions. Wheat has the potential to contribute in the mitigation of this problem as it already provides 20% calories and 20% protein in the human diet and is a staple food in many parts of the world (Reynolds *et al.*, 2012). Because of its adaptation to a wide range to climatic conditions, bread wheat is the most widely grown cereal in the world making it an easily available food commodity. However, to meet the future demand wheat production needs to increase by 2.4% as compared to current gains of 0.9% per year (Fig. 1.1) (Ray *et al.*, 2013).

Need of *Ae. tauschii* for wheat improvement

Wheat is an allopolyploid species, which received its sub-genomes from three distinct diploid species; sub-genome A from *Triticum urartu* (AA), sub-genome B from a close relative of extant *Aegilops speltoides* (SS), and sub-genome D from *Ae. tauschii* (DD) (Salamini *et al.*, 2002; Petersen *et al.*, 2006). Firstly, A and B sub-genome donors hybridized to produce the tetraploid *Triticum turgidum*, which then hybridized with diploid *Ae. tauschii* to produce

hexaploid wheat. Two polyploidization events coupled with rapid domestication of tetraploid wheats, resulted in the reduction in genetic diversity in hexaploid wheat compared to its wild progenitors (Kam-Morgan *et al.*, 1989; Lubbers *et al.* 1991; Akhunov *et al.*, 2010). Genetic diversity bottlenecks due to small number of founder lines as a result of domestication (Tanksley & McCouch, 1997; Flint-Garcia *et al*., 2013), and ploidy change (Iqbal *et al*., 2001) have been observed in many crop species, such as maize and cotton.

Despite this limited diversity, plant breeders have made substantial progress in developing wheat cultivars that are high yielding, disease and insect resistant, and are resilient to abiotic stresses. However, the practice of artificial selection has furthered the loss of genetic diversity in wheat (Fig. 1.2), and has created a bottleneck for the genetic diversity. Consequently, only a handful of alleles are represented in the elite germplasm. As a component of the rate of genetic gain, genetic diversity is the cornerstone for making progress in plant breeding for polygenic traits, such as yield, and tolerance to abiotic stresses. Therefore, increasing the genetic diversity for wheat improvement is a part of addressing the challenges to surpass the 1% gain per year in yield and reach the needed 2.4% per year (Fig. 1.1).

Genetic diversity in *Ae. tauschii*

Ae. tauschii is split into two major sub-populations, called Lineage1 (L1) and Lineage2 (L2), of which, L1 is generally ssp. *tauschii* type, and L2 is ssp. *strangulata* type (Lubbers *et al.*, 1991; Dvorak *et al.*, 1998). L2 is the presumed donor of D sub-genome of hexaploid wheat (Jaaska, 1978; Nakai, 1979; Nishikawa *et al.*, 1980). L1 and L2 were further split into two sub-groups along the longitudinal and altitudinal gradients, respectively (discussed in Chapter 3).

Ae. tauschii carries resistance to diseases and insects, such as leaf rust, stem rust, and Hessian fly (Gill, 1986; Cox *et al*., 1992; Rouse *et al*., 2011). However, much more genetic

diversity is present that can be used to improve wheat crop for abiotic stresses, and quality traits. Based on the analysis of phenotypic data and geographical distribution, we found that most of the disease resistance is confined to the center of origin, around the Caspian Sea (Fig. 1.3). However, insect resistance is spread across the natural habitat including central Asian countries (Fig. 1.3).

Comparison of wheat sub-genomes and *Ae. tauschii* revealed that the sub-genome D is genetically least diverse compared to its counterparts and the wild progenitor (Akhunov *et al.*, 2010). This is possibly because in nature only one or a few *Ae. tauschii* accessions were involved in the origin of wheat (Lagudah *et al.*, 1991). This presents an opportunity to use this untapped genetic diversity for wheat improvement. Due to the crossing difficulty, *Ae. tauschii* has been utilized sparingly in wheat breeding via direct crossing (Gill & Raupp, 1987), but mostly via bridge crossing that involves generating a synthetic hexaploid wheat by crossing a tetraploid wheat with goat grass (McFadden & Sears, 1946). Synthetic wheat is then used to introgress wild genetic diversity in wheat, however, it incorporates the genetic diversity in all three genomes, resulting in overall reduced genetic diversity in the D sub-genome relative to A and B sub-genomes. Therefore, broadening the genetic base of the D sub-genome will present further opportunities for wheat improvement.

Efficient curation of the genebanks

Genebanks play an important role as a platform for conserving and distribution of germplasm. More than 1700 genebanks around the world are holding over 7 million accessions of different plant species (Singh *et al*., 2012). However, after curating *Ae. tauschii* accessions in three different genebanks—Wheat genetics resource center (WGRC), International Maize and Wheat Improvement Center (CIMMYT), Punjab Agricultural University (PAU)—we found over

50% duplicated accessions across the genebanks (discussed in Chapter 2). If this holds true for other genebanks, and other species too, we are vastly overestimating the genetic diversity in terms of number of accessions that are present in the genebanks. We have developed a protocol to identify the redundant accessions, which can be used to rapidly curate current collections, and will facilitate the identification of duplications in the future collection endeavors. Identifying duplicated accessions will facilitate their efficient use in wheat breeding and improvement for desirable traits.

Utilizing *Ae. tauschii* genetic diversity

Presence of genetic diversity is certainly a boon for breeders and geneticists, however it can be difficult to choose a limited number of accessions to focus effort for gene introgression. Historically, the wild accessions have been selected based on their phenotypic performance under a specific environment, which is an unreliable method because accessions might not perform similarly in a different environment. Accessions carrying important alleles but exhibiting overall poor phenotype, would never be utilized based on their phenotypic data. Therefore, the application of genomic tools to assess and select the most genetically diverse accessions is more realistic. We have selected a MiniCore consisting of 40 *Ae. tauschii* accessions that captures more than 95% of the allelic diversity in whole collection (discussed in Chapter 3). Utilizing MiniCore will help breeders capture majority of the novel genetic diversity present in the whole collection by significantly reducing the size of working accessions by several fold.

Mapping one new and three known Hessian fly resistance genes in winter wheat background

Hessian fly (HF; *Mayetiola destructor* Say) is an important insect pest of wheat worldwide. Thirty-four resistance genes have been identified, however, rapid emergence of new HF biotypes necessitates the discovery of new resistance genes (Tan *et al.*, 2017). With that goal in mind, a new gene, *H2147*, providing resistance to Hessian fly biotype GP (Great Plains) was introgressed from an *Ae. tauschii* accession, *KU2147* (discussed in Chapter 4). Additionally, three previously identified genes—*H5*, *H10* and *H13*—were mapped using genotyping-by-sequencing (GBS). All genes were mapped on single chromosomal positions. Mapping the genes with linked marker information will facilitate their use in insect resistance wheat breeding, and will provide an effective and sustainable approach to control HF.

Conclusions

Wheat is an important cereal that is suffering from reduced genetic diversity due to evolutionary bottleneck and selective breeding practices. However, wheat wild relatives, such as *Ae. tauschii*, can be used to mitigate this problem. Selecting a handful of wild individuals can be a daunting task, but genomic tools can help in selecting highly diverse accessions. Out of a total collection of 574 accession, just forty such accessions were selected that capture more than 95% allelic diversity, and could be used to broaden the genetic base of wheat D sub-genome. Utilization of these accessions will facilitate the use of untapped genetic diversity present in *Ae. tauschii*, and help breeders develop disease and insect resistant, and climate resilient wheat varieties to strengthen the food security.

References

Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, Coleman-Derr D, Conley EJ, Crossman CC, *et al.* 2010. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**: 702.

T S Cox, D L Wilson, B S Gill, S Leath, W W Bockus, L E Browder WJR. 1992. Resistance to foliar diseases in a collection of *Triticum* tauschii germ plasm. *Plant Disease*: 1061.

Dvorak J, Luo MC, Yang ZL, Zhang HB. 1998. The structure of the *Ae. tauschii* genepool and the evolution of hexaploid wheat. *Theoretical and Applied Genetics* **97**(4): 657-670.

Flint-Garcia SA. 2013. Genetics and consequences of crop domestication. *Journal of agricultural and food chemistry* **61**:8267-8276.

Gill BS. 1986. Resistance in *Aegilops* squarrosa to Wheat Leaf Rust, Wheat Powdery Mildew, Greenbug, and Hessian Fly. *Plant Disease* **70**(6): 553-553.

Gill BS, Raupp WJ. 1987. Direct Genetic Transfers from *Aegilops*-Squarrosa L to Hexaploid Wheat. *Crop Science* **27**(3): 445-450.

Iqbal MJ, Reddy OU, El-Zik KM, Pepper AE. 2001. A genetic bottleneck in the'evolution under domestication'of upland cotton Gossypium hirsutum L. examined using DNA fingerprinting. *Theoretical and Applied Genetics* 103(4): 547-554.

Jaaska V. 1978. NADP-dependent aromatic alcohol dehydrogenase in polyploid wheats and their diploid relatives. On the origin and phylogeny of polyploid wheats. *Theoretical and Applied Genetics* **53**(5): 209-217.

Kam-Morgan L, Gill B, Muthukrishnan S. 1989. DNA restriction fragment length polymorphisms: a strategy for genetic mapping of D genome of wheat. *Genome* **32**(4): 724-732.

Kihara H. 1944. Discovery of the DD-analyser, one of the ancestors of *Triticum* vulgare. *Agric Hortic* **19**: 13-14.

Lagudah ES, Appels R, Brown AHD, Mcneil D. 1991. The Molecular-Genetic Analysis of *Triticum*-Tauschii, the D-Genome Donor to Hexaploid Wheat. *Genome* **34**(3): 375-386.

Lev-Yadun S, Gopher A, Abbo S. 2000. Archaeology. The cradle of agriculture. *Science* **288**(5471): 1602-1603.

Lubbers EL, Gill KS, Cox TS, Gill BS. 1991. Variation of Molecular Markers among Geographically Diverse Accessions of *Triticum*-Tauschii. *Genome* **34**(3): 354-361.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, International Wheat Genome Sequencing C, Jakobsen KS, Wulff BB, *et al.* 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**(6194): 1250092.

McFadden FE, Sears ER. 1946. The origin of *Triticum* spelta and its free-threshing hexaploid relatives. *J Hered* **37**: 81 107.

Nakai Y. 1979. Isozyme variations in *Aegilops* and *Triticum*, IV. The origin of the common wheats revealed from the study on esterase isozymes in synthesized hexaploid wheats. *The Japanese Journal of Genetics* **54**(3): 175-189.

Nishikawa K, Furuta Y, Wada T. 1980. Genetic studies on alphaamylase isozymes in wheat. III. Intraspecific variation in *Aegilops* squarrosa and birthplace of hexaploid wheat. *The Japanese Journal of Genetics* **55**(5): 325-336.

Petersen G, Seberg O, Yde M, Berthelsen K. 2006. Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Mol Phylogenet Evol* **39**(1): 70-82.

Ray DK, Mueller ND, West PC, Foley JA. 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PLoS One* **8**(6): e66428.

Reynolds M, Foulkes J, Furbank R, Griffiths S, King J, Murchie E, Parry M, Slafer G. 2012. Achieving yield gains in wheat. *Plant Cell Environ* **35**(10): 1799-1823.

Rouse MN, Olson EL, Gill BS, Pumphrey MO, Jin Y. 2011. Stem rust resistance in *Ae. tauschii* germplasm. *Crop Sci* **51**:2074–2078.

Salamini F, Ozkan H, Brandolini A, Schafer-Pregl R, Martin W. 2002. Genetics and geography of wild cereal domestication in the near east. *Nat Rev Genet* **3**(6): 429-441.

Singh AK, Varaprasad K, Venkateswaran K. 2012. Conservation Costs of Plant Genetic Resources for Food and Agriculture: Seed Genebanks. *Agric Research* **1**(3): 223-239.

Tanksley SD, McCouch SR. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**(5329): 1063-1066.

Tan CT, Yu H, Yang Y, Xu X, Chen M, Rudd JC, Xue Q, Ibrahim AMH, Garza L, Wang S, *et al.* 2017. Development and validation of KASP markers for the greenbug resistance gene Gb7 and the Hessian fly resistance gene H32 in wheat. *Theor Appl Genet*: 1-18.

United Nations, Department of Economic and Social Affairs, Population Division. 2015. World Population Prospects: The 2015 Revision, Key Findings and Advance Tables.

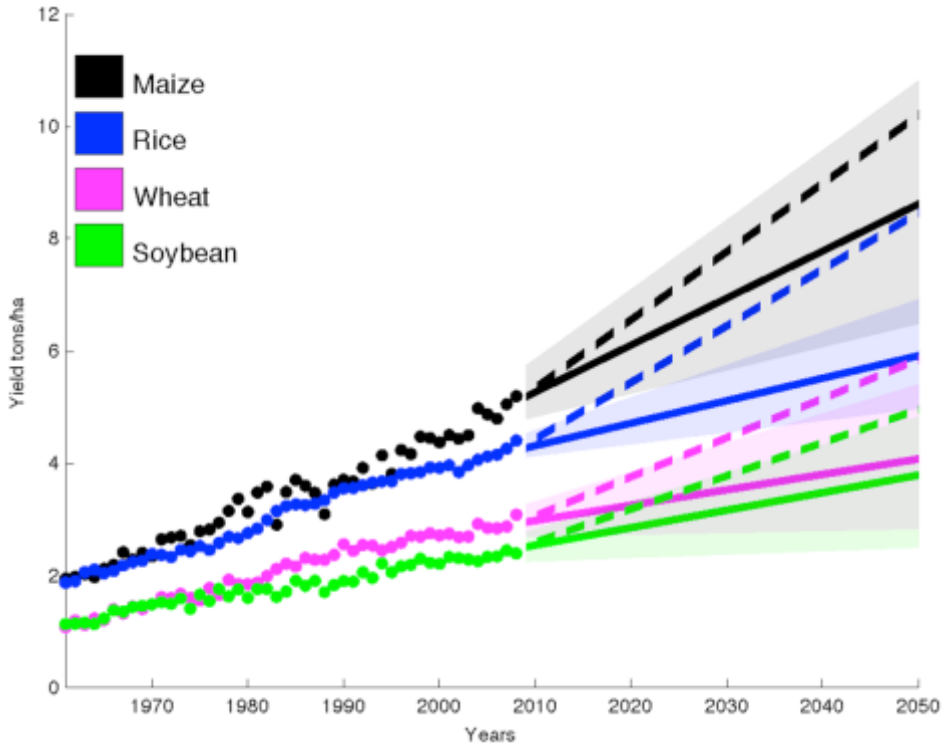

Figure 1.1. Yield trends for maize, rice, wheat, and soybean, as described in Ray *et al.* (2013).

Solid dots represent the observed data for each year up to year 2008. Solid lines from 2009 represent projected yield for each crop up to 2050 with current trends. Dotted lines represent the desired yield increase trends to double the food production by 2050. Shaded area represents 90% confidence interval from 99 bootstrap samples.

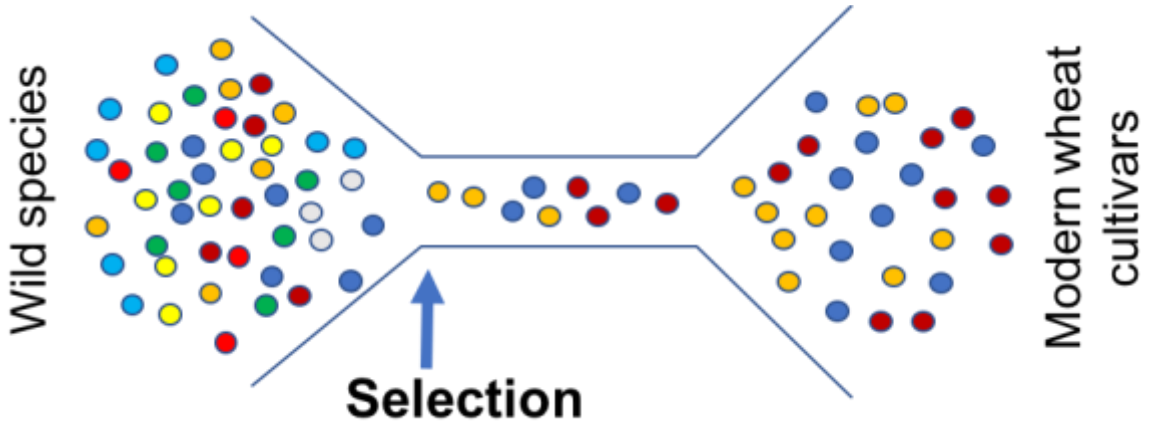

Figure 1.2. An illustration of reduction in genetic diversity of modern wheat cultivars.

Range of colored circles on the left side of the pipeline represents allelic diversity present in wild species. When the artificial selection or domestication (depicted by narrow passage) is performed, only a handful of alleles are selected, followed by the boom in population sizes resulting in only the presence of selected alleles. The loss of many colored circles in modern wheat cultivars represent the loss of genetic diversity.

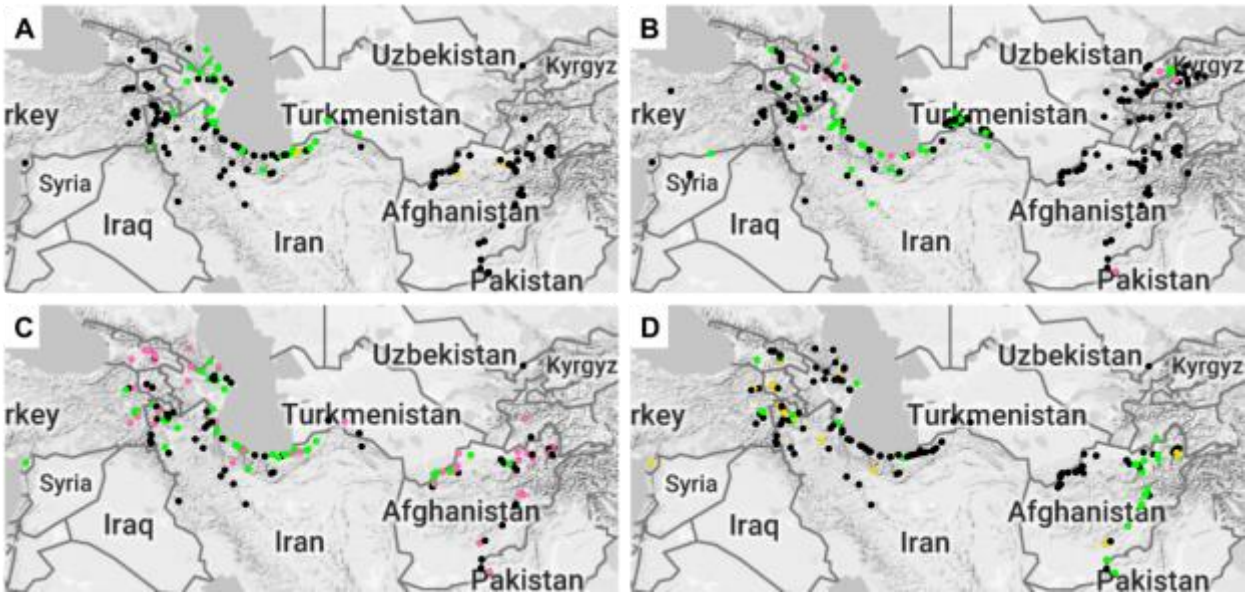

Figure 1.3. Geographical distribution of the known phenotypes for *Ae. tauschii* accessions for (A) Leaf rust, (B) Stem rust, (C) Hessian fly, and (D) Wheat curl mite.

Green dots represent resistant accessions, gold represents moderately resistant, pink represents segregating accessions, black represent no phenotypic data available. Evident from this distribution is that the resistance for fungal pathogens is only present in the center of origin for wheat around Caspian Sea, but it is spread across the natural habitat for insect pests.

Chapter 2 - Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions

Abstract

Genebanks are valuable resources for crop improvement through the acquisition, *ex-situ* conservation and sharing of unique germplasm among plant breeders and geneticists. With over seven million existing accessions and increasing storage demands and costs, genebanks need efficient characterization and curation to make them more accessible and usable and to reduce operating costs, so that the crop improvement community can most effectively leverage this vast resource of untapped novel genetic diversity. However, the sharing and inconsistent documentation of germplasm often results in unintentionally duplicated collections with poor characterization and many identical accessions that can be hard or impossible to identify without passport information and unmatched accession identifiers. Here we demonstrate the use of genotypic information from these accessions using a cost effective next generation sequencing platform to find and remove duplications. We identify and characterize over 50% duplicated accessions both within and across genebank collections of *Aegilops tauschii*, an important wild relative of wheat and source of genetic diversity for wheat improvement. We present a pipeline to identify and remove identical accessions within and among genebanks and curate globally unique accessions. We also show how this approach can also be applied to future collection efforts to avoid the accumulation of identical material. When coordinated across global genebanks, this approach will ultimately allow for cost effective and efficient management of germplasm and better stewarding of these valuable resources.

Introduction

With an estimate of more than 1 billion underfed people in the world (Martínez-Martínez & Calvo, 2010; FAO *et al.*, 2014) and projected human population growth to over 9 billion by 2050 (United Nations *et al.*, 2015), there is increased food insecurity risk and an even a greater challenge to global food supply. To meet the future demand food production needs to be doubled (FAO, 2009; Tilman *et al.*, 2011) in the midst of shrinking resources (Ray *et al.*, 2013). A critical raw ingredient for continued crop improvement is genetic diversity. Although there is tremendous diversity among flowering plants, humans cultivate only a handful of them for food and feed, with about 90% of the food and feed coming from only ten cultivated crop species (Tanksley & McCouch, 1997) (Gruissem *et al.*, 2012). Great opportunities exist to domesticate new plant species and improve the existing crop plants (DeHaan *et al.*, 2016). Genetic diversity present in wild crop relatives and conserved in genebanks are a source of novel genes that increase yield, resistance to pests and disease and abiotic stress.

Genebanks play an imperative role in *ex-situ* germplasm conservation that is critical for crop improvement. These facilities provide infrastructure for storage, a platform for sharing, and opportunity for better access and utilization of the germplasm. More than 1700 genebanks around the world stock over 7 million plant accessions (Singh *et al.*, 2012), of which only a small number are characterized, and few are ever used for crop improvement (McCouch *et al.*, 2013). Although genebanks are crucial for aforementioned reasons, they are expensive to establish and manage (McCouch *et al.*, 2013). Therefore, to maximize the value of this investment and of the germplasm resources, strategies for efficient genebank management are needed.

Researchers have implemented different strategies to prioritize a limited number of potentially useful accessions from genebanks that can be used for crop improvement. These

strategies include selecting accessions based on their phenotype and associated passport data. One example of such strategies is Focused Identification of Germplasm Strategy (FIGS) that works on the premise that the adaptive traits shown by the accessions is the direct result of environmental conditions of their respective place of origin, and the genetic diversity can be maximized by sampling accessions based on their diverse contrasting geographic regions (Bari *et al.*, 2012; Khazaei *et al.*, 2013). However, accessions stored in the genebanks are often missing the phenotypic and passport data, or could be associated with incorrect passport data, which limits the application of FIGS. Other limitations of such strategies include the high cost of phenotyping and limited resources such as space and manpower to do such screening on a larger scale. Therefore, cheaper and reliable methods that are free from these kinds of uncertainties are needed.

Contrary to the unreliable phenotypic and passport information, genotypic characterization of accessions should provide better curation of genebanks and optimize the use of genetic diversity. Modern tools and techniques such as next-generation sequencing (NGS) and genotyping-by-sequencing (GBS) can be used to rapidly and cost-effectively characterize germplasm stored in genebanks (Poland & Rife, 2012). Data generated by this approach can be used for identifying identical accessions (duplications) within and among genebanks, characterizing genomic diversity (Huang *et al.*, 2014), inferring population structure (Elshire *et al.*, 2011) and imputing missing passport information. Identifying and removing identical accessions from genebanks reduces the cost while increasing the efficiency of managing and utilizing genebank resources.

Consortiums such as the DivSeek initiative (http://www.divseek.org) exist with a vested interest in genotyping the germplasm stored in genebanks for the purpose of genetically

characterizing these resources and optimizing the use of the genetic diversity. The Wheat Genetics Resource Center (WGRC; http://www.k-state.edu/wheat-iucrc), an NSF Industry/University Cooperative Research Center, located at Kansas State University in Manhattan, KS, USA, is another example of such effort to characterize wild species stored in the in-house and collaborative genebanks. WGRC primarily specializes as a working collection of wheat genetic diversity and focuses on collecting, evaluating, identifying and mobilizing the genetic diversity. Other major genebanks are managed by the Consultative Group on International Agriculture Research (CGIAR) center throughout the world such as the International Maize and Wheat Improvement Center (CIMMYT; Mexico). CIMMYT holds over 105,000 Triticeae accessions in their global genebank outside of Mexico City. Another important CGIAR genebank with over 41,000 Triticeae accessions at the International Center for Agriculture Research in Dry Areas (ICARDA) housed in Aleppo, Syria has been lost from the turmoil in that region (https://www.genesys-pgr.org). This further highlights the need to understand the status of shared and duplicated accessions within and across genebanks. In addition, there are numerous national genebanks throughout the world such at the Punjab Agricultural University (PAU; Ludhiana, India) where accessions of local importance are stored and utilized for germplasm improvement and breeding.

Modern hexaploid bread wheat (*Triticum aestivum* L.) is a critical focus to mitigate the upcoming food security challenge in coming decades. In the context of continued wheat improvement through breeding, maintaining and increasing genetic diversity in wheat is very important. Due to genetic bottlenecks from domestication and modern breeding, wheat has a limited genetic base. Its domestication coexisted with the advent of agriculture about 10,000 years ago (Renfrew, 1973; reviewed in Bell, 1987; Lev-Yadun *et al.*, 2000; Marcussen *et al.*,

2014). Three distinct diploid species—*Triticum urartu* (AA), a relative of the extant *Aegilops speltoides* (BB), and *Aegilops tauschii* (DD)—contributed to the origin and evolution of polyploid wheat (AABBDD). First natural hybridization of *Triticum urartu* and B-genome donor resulted in tetraploid *Triticum turgidum* (AABB) wheat around 0.58-0.82 million years ago (Jordan *et al.*, 2015) followed by a second whole-genome hybridization with *Ae. tauschii* (DD) (Kihara, 1944; McFadden & Sears, 1946) in the fertile crescent around the Caspian Sea, to give rise to modern hexaploid wheat. The limited hybridization with *Ae. tauschii* due to change in ploidy, followed by domestication and improvement has severely limited the genetic diversity of the wheat D genome (Akhunov *et al.*, 2010). The presence of great genetic diversity in these wild relatives provides an excellent resource for continued improvement.

As a proof of concept for genebank curation, we used *Ae. tauschii* as a model for this study while providing valuable and needed curation of several important repositories for this species. The main objectives of this study were to (i) genotype the entire collections of *Ae. tauschii* from three different genebanks using a cost effective and robust reduced representation sequencing, (ii) identify identical accessions within genebanks using genotypic data, (iii) identify identical accessions between genebanks using genotypic data, and (iv) develop protocols for efficiently curating genebanks.

Methods

Germplasm acquisition

A total of 1143 accessions of *Ae. tauschii* were assessed, which included 568 accessions from the Wheat Genetics Resource Center (WGRC, Kansas State University), 187 accessions from Punjab Agricultural University (PAU; Ludhiana, India), and 388 accessions from Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT; Mexico) (Supplementary Table

B.1). The germplasm consisted of accessions collected from natural habitat (Fig. 2.1) and accessions received from other genebanks.

DNA extraction and Genotyping

Two approaches for DNA extraction and the GBS libraries preparation were implemented for WGRC and PAU accessions (hereafter referred to as Set 1), and CIMMYT accessions (hereafter referred to as Set 2). For Set 1, young leaf tissues from single 2-3 weeks old seedlings were collected in 96 well plates. Genomic DNA was extracted using Qiagen BioSprint 96 DNA Plant Kit (QIAGEN, Hilden, Germany) and quantified with Quant-iT™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA). At least one random well per plate was left blank with known position for quality control and library integrity. GBS libraries were prepared following the protocol from Poland *et al.* (2012). Briefly, the libraries were prepared in 95-plex using 384A adapter set. For complexity reduction, DNA for each sample was digested using two enzymes – rare cutter *PstI* (CTGCAG), to which the uniquely barcoded adaptors were ligated, and frequent cutter *MspI* (CCGG), to which the common reverse adapter was ligated. All samples from a single plate were pooled and amplified using polymerase chain reaction (PCR). Detailed protocol can be found on Wheat Genetics and Germplasm Improvement website (http://wheatgenetics.org/download/category/3-protocols). Libraries were sequenced on ten lanes in total on Illumina HiSeq2000 (Illumina, San Diego, CA, USA) platform at University of Missouri (UMC; Columbia, Missouri) and McGill Univesity-Génome Quebec Innovation Centre (Montreal, Canada) facility. To compute the error rate for the GBS, 76 WGRC accessions were randomly chosen, and were sequenced as biological replications (different seedlings) using the abovementioned protocol.

For Set 2, *Ae. tauschii* accessions were planted in greenhouse in plots. Leaves from single seedling plants were taken and DNA was extracted using modified CTAB (cetyltrimethylammonium bromide) method (Hoisington, 1992) and quantified using NanoDrop spectrophotometer V2.1.0 (ThermoFisher Scientific, Waltham, MA, USA). Genotyping was performed at DArT, Canberra, Australia (http://www.diversityarrays.com) using DArTseq (Li *et al.*, 2015) methodology that has been used in recent years at CIMMYT (Li *et al.*, 2015; Sehgal *et al.*, 2015; Vikram *et al.*, 2016). DArTseq is a combination of diversity array technology (DArT) (Jaccoud *et al.*, 2001; Wenzl *et al.*, 2004) complexity reduction and next-generation sequencing (NGS) methods. Two optimized enzyme sets, *PstI-HpaII* and *PstI-HhaI*, were used for complexity reduction. Samples were sequenced twice using two different 4bp cutters on one end of the RE fragments (*HpaII* and *HhaI*) on a total of nine lanes.

SNP discovery

Single nucleotide polymorphisms (SNPs) were discovered and typed with TASSEL-GBS (Glaubitz *et al.*, 2014) framework (http://www.maizegenetics.net) using an in-house written Java plugin and a modified Java pipeline without reference genome. In brief, 64bp long valid tags (containing restriction cut site and a barcode) were extracted from each sample, and then similar tags (up to 3bp differences) were internally aligned to find SNPs. To test putative tag pairs for allelic SNP calls, Fisher exact test was performed on all aligned tag pairs with one to three nucleotide differences. Tag pairs that failed the test at $P \leq 0.001$ were considered biallelic and converted to SNP calls (Poland, J *et al.*, 2012). As the accessions are inbred lines, this test determined allelic tags that are disassociated (e.g. only one of the two alternate tags present in any given individual) and can be considered alternate tags for SNP alleles at the same locus. Due to the differences in library preparation for Set 1 and Set 2, the tag discovery step was performed

using only Set 1 accessions, and then the discovered tags were used as reference to produce SNPs for both sets.

Statistical analyses, allele matching and error computation

Data analyses and genotype curation were performed using custom scripts in *R* statistical language (R Core Team, 2015) to find identical accessions within and among genebanks. In addition to hierarchical clustering (Supplementary Fig. B.1), an identity matrix was computed by pairwise comparison of accessions across all SNP sites. Hierarchical clustering group individuals based on the relative genetic distance between individuals, whereas, pairwise allele matching provides an absolute percent identity by state (IBS) coefficient between all individuals. Although, clustering can provide an independent support for allele matching, it is hard to interpret clustering to identify identical accessions. However, clustering can provide a quick method to identify obvious outliers and misclassified accessions (Supplementary Fig. B.1). For clustering, population-level SNP filtering was performed to retain the SNPs with ≤50% missing data. In contrast, for pairwise comparison, only those SNP sites without missing data and homozygous in both individuals were used for comparison. A stringent threshold of 99% identity was used to consider two accessions the same to account for a 1% sequencing and alignment error rate. Accessions with ≥99% identity were considered identical within and/or across genebanks. Percent Identity by State (pIBS) was computed using the following equation I:

$$pIBS_{ij} = \frac{\sum_{x=1}^{n}(allele_{ix} == allele_{jx})}{n}$$

where, $pIBS_{ij}$ is the percent Identity by State for a given pair of accessions *i* and *j*, $allele_{ix}$ and $allele_{jx}$ are the x^{th} allele of accessions *i* and *j*, respectively, == sign represents an exact successful match (identity by state) between two alleles, and *n* is the total number of SNP sites in a pairwise comparison. The same equation was used to compute pIBS for an accession with its biological

rep for error rate computation. In that case *i* and *j* represents the original accession and its biological replicate, respectively. Accessions with pIBS ≥99% (0.99) were grouped together in an arbitrary group number. Group size was computed as number of accessions in a group.

An error rate was computed using biological replicates for 76 accessions. Single to multiple seeds were grown for each accession, DNA was extracted and sequencing performed as explained above. The error rate was computed using the following equation II:

$$error\ rate = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{m}(1 - pIBS_{ij})}{m}$$

where, *n* is the number of accessions with biological replicates, $pIBS_{ij}$ is the percent IBS for *i*th accession with its *j*th replicate, and *m* is the number of replicates for a given accession.

Gliadin Profiling

To complement our GBS identity results, we extracted and profiled gliadin proteins from five independent groups of identical accessions that were found to be the same with GBS (Supplementary Table B.2). A single seed per accession was crushed in pestle and mortar to fine flour and mixed with 70% ethyl alcohol and stored at -4 ºC for 24 hours. Following the protein extraction, samples were prepared using Bio-Rad Experion Pro260 kit (Bio-Rad, Hercules, California) following manufacturer's instructions, and loaded on to an Experion Pro260 chip. The chips were read using Bio-Rad Experion automated electrophoresis system (Bio-Rad, Hercules, California). Virtual gel images were analyzed to compare accessions for identical protein banding patterns. For later comparison of protein profiling and GBS for two samples, multiple seeds were subjected to both procedures, where half of the seed was used for protein extraction and the other half with intact embryo was used for germination and tissue collection for DNA extraction.

Imputing passport information

To facilitate the reduction of missing data and better curation of genebanks, we used genomic data and STRUCTURE (Pritchard *et al.*, 2000) software to impute the missing passport information for 26 WGRC accessions. For imputation, all the accessions with available passport information were used as learning samples and the remaining with missing to be imputed. The STRUCTURE parameters were set as follows: 10,000 burn-in iterations followed by 10,000 MCMC iterations, POPDATA=1, USEPOPINFO=1, GENSBACK=1, LOCIPOP=1, and all other parameters left at default settings. This resulted in posterior probabilities for each accession belonging to a specific geographical group with certain probability.

Results

Sequencing and SNP genotyping

GBS generated ~2 billion 100bp reads for Set 1, and DArTSeq generated ~1 billion 77bp reads for Set 2, of which, 1.6 billion (83.4%) in Set 1 and 861 million (85.4%) contained expected sample barcodes followed by a restriction site. On average, each sample generated 1.9 million and 1.4 million barcoded reads for Set 1 and Set 2, respectively. Using these reads, discovery step in TASSEL-GBS pipeline found a total of ~93 million unique 64bp tags. Each accession contributed an average of 81,365 unique tags that were aligned internally to find putative SNP sites, which resulted in 91,545 SNPs. Proportion of missing SNP data ranged from 0.6% to 78.9%. Population-level SNP filtering with ≤50% missing data, retained 29,555 SNPs that were used for cluster analysis. For pIBS, 20,844 pairwise comparisons were performed on average between any two accessions.

Clustering and identifying identical accessions

Two different analyses were performed to identify identical accessions; a cluster analysis and allele matching. Cluster analysis (Supplementary Fig. B.1) provides a quick method to cluster accessions based on genetic distances, however it cannot find identical accessions *per se*. For curating genebanks, cluster analysis should be used as a first step to group phenotypically cryptic accessions outside of the species under study and identify other outliers. From the cluster analysis, we observed the strong population structure between lineage 1 and lineage 2 that is known and previously reported in in *Ae. tauschii* (Dvorak *et al.*, 1998). As expected, we could assign all accessions into two large clusters, and identified three outliers which were removed from subsequent analysis (Supplementary Fig. B.1). Accession TA3429 was found to be an outlier in STRUCTURE analysis. Two other accessions, one each from PAU and CIMMYT, clustered with TA3429 to form an outlier group. Corroborated by allele matching analysis, these outliers did not match with any other accession, supporting evidence that they have been misidentified as *Ae. tauschii*.

Contrary to cluster analysis, allele matching provides an absolute percent IBS coefficient that can be used to identify identical accessions. Based on allele matching, different accessions had pairwise identity ranging from 37.5-99.9% (Supplementary Fig. B.2). Each genebank resulted in a bimodal distribution of pIBS because of the strong population structure within *Ae. tauschii*. The higher pIBS peak represents the percent identity within subpopulations, and lower pIBS peak represents between subpopulations. With genotyping error, it is not possible to expect a 100% allelic identity for accession that should be considered the same. For this study, we implemented 99% allelic identity threshold for declaring accessions identical. This was initially based on expected sequencing error rates and confirmed with biological sample replicates.

Minimum and maximum number of duplicated accessions were found in WGRC (25.88%) and PAU (54.01%), respectively, with CIMMYT having 43.04% duplicated accessions (Fig. 2.2). Combined across all genebanks, about 50% accessions were putatively duplicated. After removing the identical accessions, the WGRC, CIMMYT and PAU had only 421 (74.12%), 221 (45.99%) and 86 (45.99%) unique accessions, respectively. Based only on these unique accessions, pairwise IBS were computed for the accessions across the genebanks. The WGRC shared 32 (12.62%) with PAU and 129 (40.19%) accessions with CIMMYT, and PAU shared 29 (18.89%) accessions with CIMMYT. Overall, all three genebanks shared 26 (10.71%) accessions (Fig. 2.3) with group size of identical accessions ranging from 2 - 44 accessions (Fig. 2.4). After grouping the accessions across all genebanks, only 564 unique accessions were found, representing over 50% duplicated accessions across the combined collections.

Error rate and efficiency

To compute the error rate of the GBS method, 76 accessions from the WGRC were resequenced and used as biological replicates. Of these 76 accessions, 11 had pIBS <99% with their respective original samples. Using the equation II, the overall error rate was computed to be 3.13%, which is higher than our 1% threshold. To investigate further, multiple seeds from these 11 accessions were planted, however, only eight accessions produced at least one plant. GBS was performed on these eight accessions as described below.

Four out of eight accessions produced only a single plant. These were resequenced and compared with their previously sequenced respective samples (original sample and biological replicate). As initially expected, all four resequenced samples matched with >99% pIBS with either the original sample or the respective biological replicates. Two of these accessions matched with their original sample and other two matched with their biological replicates. These

results point to the possibility of sample contamination that resulted in bad GBS data in one of the two initial GBS runs. Another possibility is that the original seed source was heterogeneous. Seed or sample mixture during the genotyping process of large number of samples is possible, however, we attempted to test the latter conjecture.

Remaining four out of eight accessions (TA1581, TA1589, TA1714 and TA2468) produced multiple plants that allowed us to test our hypothesis that the original seed source was heterogeneous. The final GBS was performed on each plant individually and compared with their respective original samples and biological replicates. TA1581 and TA1589 matched nicely with their original sample and all other replicates within this GBS run, but not the previous biological replicate. This points to the possibility that the sample contamination might have happened during the sequencing of previous biological replicates for these two accessions. In contrast, resequenced samples for TA2468 matched with >99% identity with the previous biological replicate and all other samples within this run, but failed to match with the original GBS. This again points to the possibility that the sample contamination might have happened during the original GBS.

For the final TA1714, a different pattern was observed. Two of the four resequenced samples matched with >99% identity with the original GBS, and the other two matched with the biological replicate. This supports our hypothesis, and presents evidence that the genebank seed source might be heterogeneous that results in lower pIBS. This is further evident in independent gliadin profiling discussed below. After removing these anomalous coefficients, the accuracy improved, and the error rate was reduced to only 0.48%, which is below our 1% threshold.

Gliadin profiling

To independently validate the GBS results, gliadin profiling was run on eight independent groups from the cluster analysis in two separate runs. Gliadin proteins were selected for independent confirmation because of their ease of extraction and polymorphic profiling pattern. The first run included ten samples from four different groups (Fig. 2.5). Per the manufacturer's manual, bands lower than 10 kD were excluded as these are system bands that are produced by the small molecules interacting with lithium dodecyl sulfate (LDS) micelles in gel-staining solution and do not carry useful information. We observed matching banding pattern for the identical samples within the groups. For the second run (Fig. 2.6), samples were included from four other different groups. As expected, the samples within all groups have similar banding pattern with the following notes. Sample TA2457 (Fig. 2.6 - Lane 7) has the similar banding pattern as other samples from Grp15 (Lanes 5 and 6) but has a smeared profile that might be due to higher amount of extracted protein. Sample TA1579 (Fig. 2.6, Lane 2) is the only accession from Grp187 and had very different banding pattern as compared to any other lane in this gel. Overall, matching banding pattern for the accessions within a group provides an independent evidence that the accession grouping based on GBS results are accurate.

Detecting accession heterogeneity

TA1714 was hypothesized to be a heterogeneous, and TA2457 a homogeneous accession based on the initial GBS grouping. To detect and confirm the heterogeneity in the source seed, these two accessions were subjected to a final GBS run. Half of the seed was crushed for protein extraction and the remaining half with intact embryo was germinated for tissue collection for GBS. For TA1714 and TA2457, 12 and 15 plants of each accession were planted, respectively, and subjected to GBS and gliadin profiling. As expected TA1714 showed heterogeneity in both

the GBS and gliadin profiling by forming two sub-groups (Fig. 2.7; red and blue box). Gliadin profiling was corroborated with GBS grouping of these samples. Contrary to TA1714, TA2457 did not show different banding pattern among individual plants from this accession (Fig. 2.8), which supports that TA2457 is homogeneous. Both gliadin profiling results match with the corresponding GBS sub-groups. Independent confirmation with gliadin profiling supports that GBS can also be implemented to detect heterogeneity in the genebank samples.

Imputing missing passport information

STRUCTURE analysis resulted in posterior probabilities ranging from 0.001-1. Higher posterior probability indicated higher likelihood that the accession belongs to a certain geographical group. Because these geographical groups are not completely isolated, we treated these groups as admixed populations, hence we used the posterior probability of 0.6 or more to assign an accession in a group. Using this analysis, we could assign 24 out of 26 accessions with missing geographical information into one of the geographic clusters. Two remaining accessions could not be assigned to any specific group because of lower probabilities (Supplementary Table B.3)

Discussion

Genotyping platform and accuracy

Selecting a genotyping platform is important when a large number of samples are of interest. We sequenced 1143 *Ae. tauschii* samples using two genotyping-by-sequencing methods. Sequence-based methods, such as GBS, are inexpensive and robust for genotyping a diverse range of uncharacterized species with complex genomes (Poland & Rife, 2012), could be combined from multiple platforms. Here, we could use newly generated GBS data for set1 and previously generated DArTSeq for set2, to find duplicated accessions and efficiently curate the

genebanks. As no prior SNP information is required for sequence based methods, they also control for ascertainment bias because the SNP discovery and genotyping is performed on the same samples. Even though GBS only captured less than 1% of the genome, it resulted in an average of 20,844 pairwise SNP comparisons for allele matching. GBS grouping complemented with gliadin profiling, a very small error rate of only 0.48% makes it is a robust tool for this type of germplasm characterization.

Collaborating with other genebanks

The ability to combine existing genotypic datasets and germplasm sharing is of great interest for genebank collaborations. As a starting point, this strategy was used on a diploid progenitor of wheat to identify unique accessions within and among genebanks. Here a coordinated effort between WGRC, CIMMYT and PAU could compare 1143 *Ae. tauschii* accessions across the genebanks and identify both identical and unique accessions across all the genebanks. Genebanks included in this study were rather smaller in size where all the accessions were genotyped and characterized, however, large scale genebanks usually lack this practice and record of duplicated accessions are often missing. Historically, even when these records are disseminated during germplasm sharing, they tend to lose track over time because of poor management practices. Therefore, the ultimate benefit of this strategy will be realized when this method is implemented globally in collaboration across all genebanks. The sequencing technology has quickly reached a point to enable globally coordinated effort among all genebanks to genetically curate these collections and find unique accessions in them. These globally unique accessions should then be prioritized and likely shared with other genebanks for additional backup of those irreplaceable accessions.

Defining globally unique accessions

We have correlated many accessions with lost or incorrect accession identifiers through genotyping these collections. Most misclassifications happen during sharing of germplasm resources between collections (Emanuelli *et al.*, 2013), which leads to significant duplication and incorrect information. Historically, germplasm was frequently shared, however, the associated metadata often was lost or misidentified, resulting in inaccurate classification and the new identifiers assigned lead to duplications in and across collections. Re-collecting at the same locations and sharing germplasm among genebanks also results in duplications within and among genebanks. We found 26-54% redundant accessions within, and a total of over 50% redundant accessions among the WGRC, CIMMYT and PAU genebanks. Our GBS results were corroborated by gliadin profiling. GBS generates genome-wide biallelic markers, whereas gliadin protein profiling samples multiple alleles from only a handful of loci, which complements and independently validating GBS results. As a starting point, we only performed this analysis for *Ae. tauschii*, but this strategy can be extended to other species with different ploidy levels stored in various genebanks. Genebanks worldwide are reported to hold over a million *Triticeae* accessions (Knüpffer, 2009). However, if our observations from this study hold true for other species, including the *Triticeae* tribe, we are vastly overestimating the number of unique germplasm accessions stored in the genebanks.

Applying genetic curation across genebanks around the world should be made a coordinated priority. Once unique accessions are identified across all collections, a globally unique ID could be generated and duplicate accessions within and between collections noted. With global curation, genebanks can better coordinate and curate collections efficiently. Currently, 482 genebanks use the GENESYS database (https://www.genesys-pgr.org) for over

3.6 million accessions which could provide a platform for establishing global curation. Such curation could also help other research endeavors, such as recently funded CGIAR Genebank Platform 2017-2022, whose main goal is to make available 750,000 accessions of crops and trees to the research community for crop improvement.

Curating passport information and metadata

Often, vital metadata associated with shared germplasm, such as geographical or species information, is missing or incorrect. Species classification is a real challenge when dealing with cryptic species. A combination of existing genomic tools and statistical analyses can be used to infer those missing pieces. We used one such combination, GBS and cluster analysis and identified outliers (Supplementary Fig. B.1). Although it is very difficult to accurately assign an accession to a geographical region at city level resolution, genotypic similarities and ancestry relationships can be used to group them together with other accession that have the metadata available. We used such methods to assign 24 out of 26 accessions to a potential geographical region of origin. Meyer (2015) noted that researchers tend to use germplasm with complete passport information and other associated metadata, which provides an incentive to collect and curate the accessions, and infer the missing information.

Future direction for germplasm collection

The role of wild germplasm in crop improvement and the need to collect and preserve as much wild diversity as possible is evident. However, a specific protocol is necessary to avoid the accumulation of redundant accessions and keep only unique ones. One such approach is presented here (Fig. 2.9). Briefly, when a new accession is collected or received, multiple seeds should be planted for tissue collection, and tissue should be collected in bulk from all plants, which was not ensured in this current study. We only sampled single seed from each accession,

and it is possible that we missed within sample heterogeneity. Genotyping should be done on the bulked tissue from several seedlings. However, because *Ae. tauschii* is a highly self-pollinated species, it is very rare to find within accession heterogeneity unless due to seed mixture. Nevertheless, if possible, multiple independent samples should be sequenced for each accession. High level of heterozygous SNP calls, and mismatches within an accession, should point to the possibility of heterogeneous seed source that can be purified using single seed descent method. Bulked genotype data should be used for comparison to an existing genotyping database to find if the new accession is unique or identical to an existing collection. If unique, a new ID should be assigned, otherwise, the accession should be grouped together with the existing group of accessions. One such case study is explained below.

Case study for collecting new accessions

About 61% of the WGRC accessions were collected in 1950s and 60s by various explorers and obtained through sharing between various genebanks. To fill the gaps in the collection sites and to preserve more genetic diversity a recent collection expedition was conducted in June 2012 by WGRC researchers. During this expedition, a total of 224 accessions of *Ae. tauschii* were collected with passport information (blue dots; Fig. 2.1). Based on our analysis, only 134 collected accessions (60%) were unique in that they did not match with any other accession, either the newly collected or the already existing accessions. Surprisingly, sixteen newly collected accessions had pIBS >99% with one or more existing WGRC accessions that were collected decades ago. More expected, 71 accessions had pIBS >99% with one or more new accessions that were from the same geographic areas. CIMMYT and PAU also had 17 and 8 matching accessions with newly collected accessions, respectively. Five accessions matched with one or more other accessions from all genebanks. Even though we collected 224 new accessions,

only about 134 (60%) were unique. These findings support implementation of a protocol for efficiently curating the genebanks in place, which is based on genotypic data.

Conclusion

There are significant costs associated with running a genebank, beginning with acquiring an accession to storing and maintaining the germplasm. Because genebanks have limited funding and resources, identifying the duplicate accessions would result in a savings on both. Cost effective genotyping methods, such as GBS, can be applied for identifying duplicate accessions, and infer missing geographical and species information. Our results indicate that we are overestimating the diversity stored in the genebanks. Ultimately, identifying unique accessions within and across the genebanks will facilitate the better use of wild germplasm, make sharing more efficient, help breeders work with genetically diverse unique individuals and make better use of the untapped genetic diversity.

Acknowledgements

This study was conducted under the auspices of the Wheat Genetics Resource Center (WGRC) Industry/University Collaborative Research Center (I/UCRC) through support of industry partners and partially funded by NSF grant contract (IIP-1338897). We would also like to acknowledge the contribution of Industry Advisory Board to improve this manuscript with their valuable comments and suggestions.

References

Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, Coleman-Derr D, Conley EJ, Crossman CC, Deal KR, *et al.* 2010. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**: 702.

Bari A, Street K, Mackay M, Endresen DTF, De Pauw E, Amri A. 2012. Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables. *Genetic Resources and Crop Evolution* **59**(7): 1465-1481.

Bell G 1987. The history of wheat cultivation. Wheat Breeding: *Springer*, 31-49.

DeHaan LR, Van Tassel DL, Anderson JA, Asselin SR, Barnes R, Baute GJ, Cattani DJ, Culman SW, Dorn KM, Hulke BS. 2016. A Pipeline Strategy for Grain Crop Domestication. *Crop Science*.

Dvorak J, Luo MC, Yang ZL, Zhang HB. 1998. The structure of the *Ae. tauschii* genepool and the evolution of hexaploid wheat. *Theoretical and Applied Genetics* **97**(4): 657-670.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**(5): e19379.

Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troggio M, Myles S, Martinez-Zapater JM, Zyprian E, Moreira FM. 2013. Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC plant biology* **13**(1): 39.

FAO. 2009. Global agriculture towards 2050. Rome.

FAO, IFAD, WFP. 2014. The State of Food Insecurity in the World 2014. Strengthening the enabling environment for food security and nutrition. Rome.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**(2): e90346.

Gruissem W, Lee CH, Oliver M, Pogson B. 2012. The Global Plant Council: Increasing the Impact of Plant Research to Meet Global Challenges. *Journal of Plant Biology* **55**(5): 343-348.

Hoisington D. 1992. *Laboratory protocols: CIMMYT applied molecular genetics laboratory*: CIMMYT.

Huang Y-F, Poland JA, Wight CP, Jackson EW, Tinker NA. 2014. Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat.

Jaccoud D, Peng K, Feinstein D, Kilian A. 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res* **29**(4): E25.

Jordan KW, Wang S, Lun Y, Gardiner LJ, MacLachlan R, Hucl P, Wiebe K, Wong D, Forrest KL, Consortium I, *et al.* 2015. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol* **16**(1): 48.

Khazaei H, Street K, Bari A, Mackay M, Stoddard FL. 2013. The FIGS (Focused Identification of Germplasm Strategy) Approach Identifies Traits Related to Drought Adaptation in Vicia faba Genetic Resources. *PLoS One* **8**(5): e63107.

Kihara H. 1944. Discovery of the DD-analyser, one of the ancestors of *Triticum* vulgare. *Agric Hortic* **19**: 13-14.

Knüpffer H 2009. Triticeae genetic resources in ex situ genebank collections. *Genetics and Genomics of the Triticeae*: Springer, 31-79.

Lev-Yadun S, Gopher A, Abbo S. 2000. The cradle of agriculture. *Science* **288**(5471): 1602-1603.

Li H, Vikram P, Singh RP, Kilian A, Carling J, Song J, Burgueno-Ferreira JA, Bhavani S, Huerta-Espino J, Payne T. 2015. A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics* **16**(1): 216.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, International Wheat Genome Sequencing C, Jakobsen KS, Wulff BB, Steuernagel B, Mayer KF, *et al.* 2014.

Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**(6194): 1250092.

Martínez-Martínez L, Calvo J. 2010. Food: The growing problem. *Nature* **466**(7306): 546-547.

McFadden FE, Sears ER. 1946. The origin of *Triticum* spelta and its free-threshing hexaploid relatives. *J Hered* **37**: 81 107.

McCouch S, Baute GJ, Bradeen J, Bramel P, Bretting PK, Buckler E, Burke JM, Charest D, Cloutier S, Cole G, *et al.* 2013. Agriculture: Feeding the future. *Nature* **499**(7456): 23-24.

Meyer RS. 2015. Encouraging metadata curation in the Diversity Seek initiative. *Nature Plants* **1**(7).

Poland J, Endelman J, Dawson J, Rutkoski J, Wu SY, Manes Y, Dreisigacker S, Crossa J, Sanchez-Villeda H, Sorrells M, *et al.* 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *Plant Genome* **5**(3): 103-113.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* **7**(2): e32253.

Poland JA, Rife TW. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. *Plant Genome* **5**(3): 92-102.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**(2): 945-959.

R Core Team. 2015. R: A language and environment for statistical computing: R Foundation for Statistical Computing.

Ray DK, Mueller ND, West PC, Foley JA. 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PLoS One* **8**(6): e66428.

Renfrew JM. 1973. Palaeoethnobotany. The prehistoric food plants of the Near East and Europe.

Sehgal D, Vikram P, Sansaloni CP, Ortiz C, Pierre CS, Payne T, Ellis M, Amri A, Petroli CD, Wenzl P, *et al.* 2015. Exploring and Mobilizing the Gene Bank Biodiversity for Wheat Improvement. *PLoS One* **10**(7): e0132112.

Singh AK, Varaprasad K, Venkateswaran K. 2012. Conservation Costs of Plant Genetic Resources for Food and Agriculture: Seed Genebanks. *Agricultural Research* **1**(3): 223-239.

Tanksley SD, McCouch SR. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* **277**(5329): 1063-1066.

Tilman D, Balzer C, Hill J, Befort BL. 2011. Global food demand and the sustainable intensification of agriculture. *Proc Natl Acad Sci U S A* **108**(50): 20260-20264.

United Nations, Department of Economic and Social Affairs, Population Division. 2015. World Population Prospects: The 2015 Revision, Key Findings and Advance Tables.

Vikram P, Franco J, Burgueño-Ferreira J, Li H, Sehgal D, Saint Pierre C, Ortiz C, Sneller C, Tattaris M, Guzman C. 2016. Unlocking the genetic diversity of Creole wheats. *Scientific reports* **6**.

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A. 2004. Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *Proc Natl Acad Sci U S A* **101**(26): 9915-9920.

Figure 2.1. Geographical distribution of the WGRC *Ae. tauschii* accessions.

Each dot represents a collection site for *Ae. tauschii* accessions. Blue dots represent newly collected accessions (June 2012), and red dots represents previously collected accessions (1950s and 60s). Two accessions from China's Shaanxi and one from Henan are not shown here to control for the size of the map.

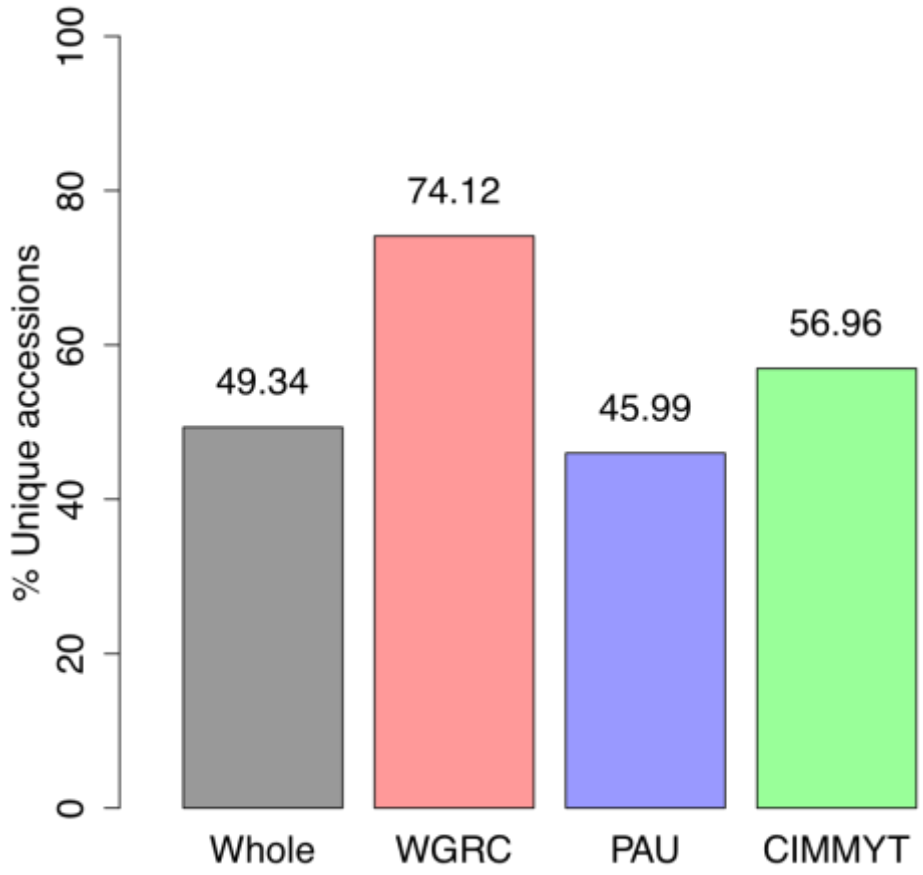

Figure 2.2. Bar plot showing percent unique accessions in whole collection, WGRC, PAU, and CIMMYT genebanks.

Values on top of each bar denotes the exact percent of unique accessions.

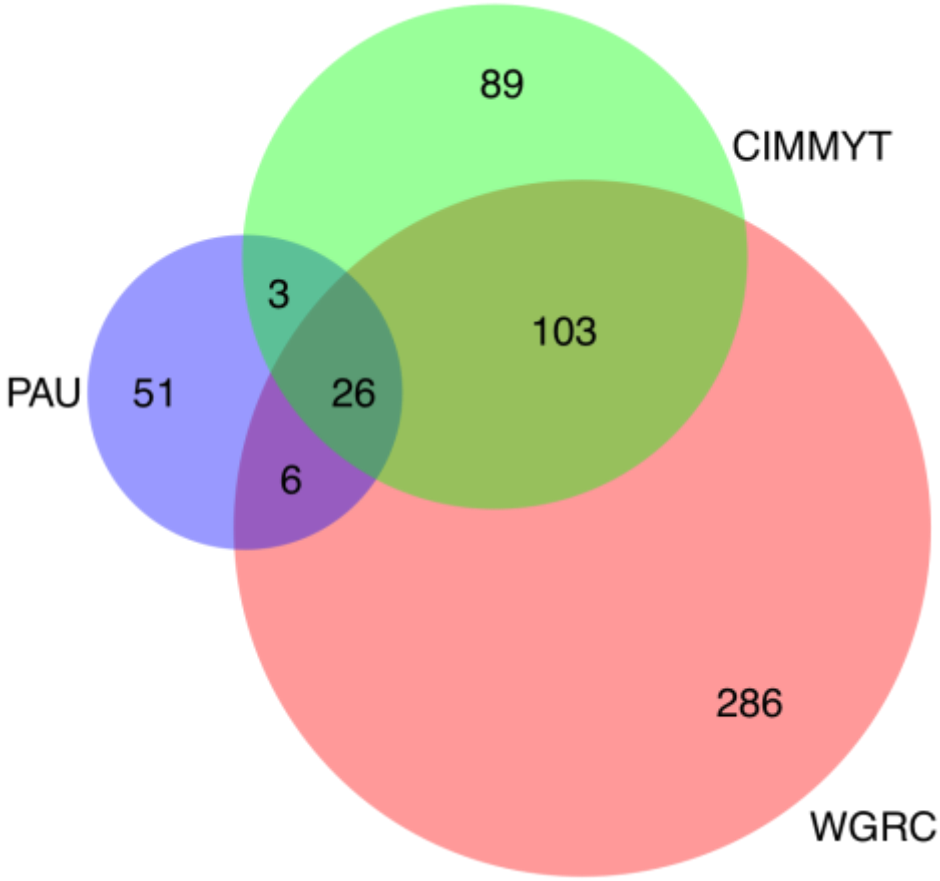

Figure 2.3. Venn diagram of shared and unique accessions among and within genebanks.

The total number for each genebank represents only unique accessions within a genebank. Numbers inside the shaded overlapping areas represent shared accessions across those specific genebanks.

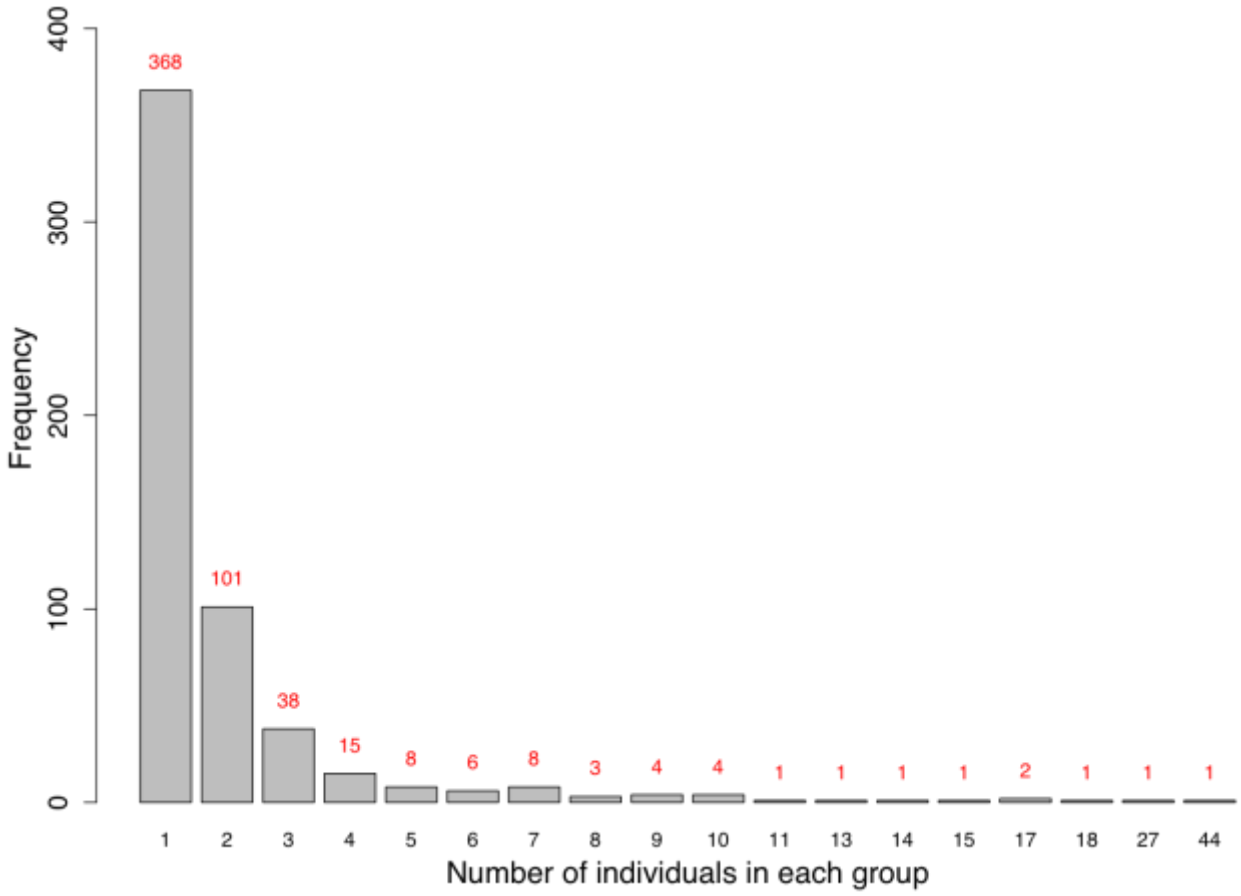

Figure 2.4. Bar plot showing frequency of each group size.

Values on top of each bar represents the exact frequency of corresponding group size listed on x-axis. Total of 368 accessions were total unique and did not match with any other accession.

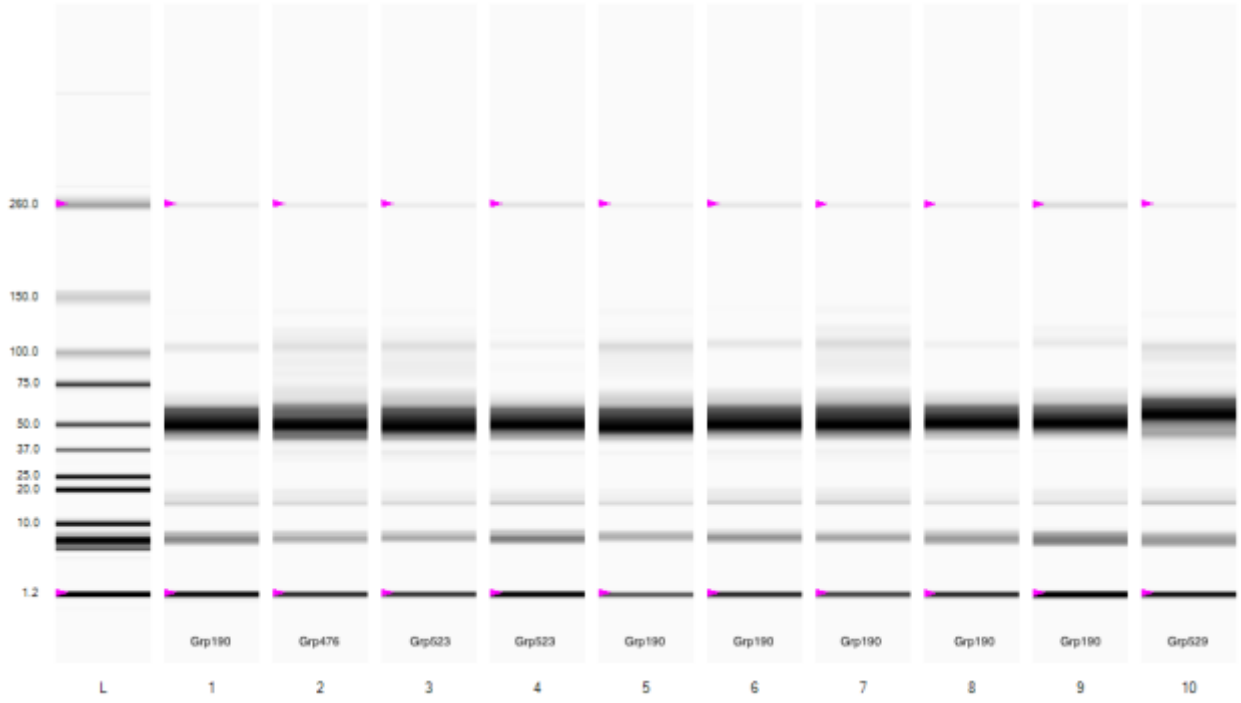

Figure 2.5. Virtual gel image of gliadin profiling for accessions from four different groups- Grp190, Grp476, Grp523, and Grp529.

Lanes 1 and 5-9 from Grp190, lane 2 from Grp476, lanes 3-4 from Grp523 and lane 10 from Grp529. As expected, lanes 2 and 10 shows different banding pattern as they are the only representative of their respective groups on this gel. Lanes 3 and 4 have similar banding pattern. Lanes 1 and 5-9 from Grp190 have similar banding pattern. This suggests that accessions within a group tend to have a similar banding pattern, which corroborates with the accession grouping with allele matching.

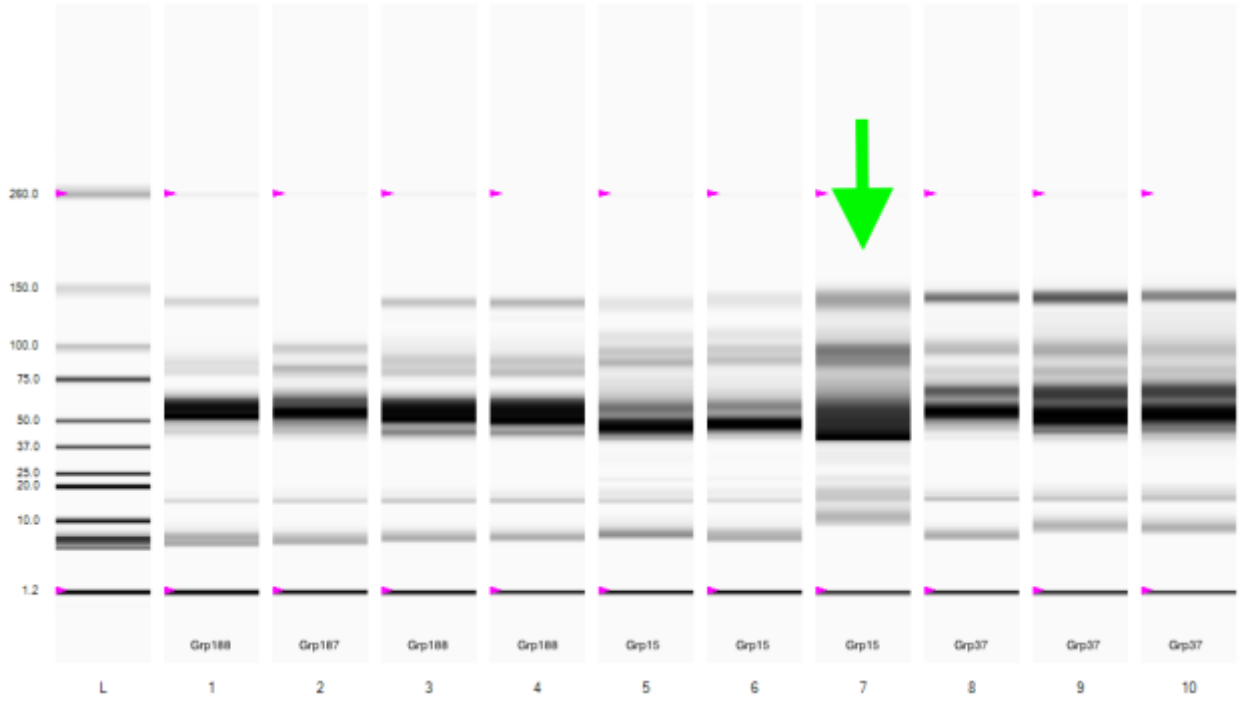

Figure 2.6. Virtual gel image of gliadin profiling for accessions from four different groups- Grp15, Grp37, Grp187, and Grp188.

Lanes 1, 3, and 4 are from Grp188; lane 2 from Grp187; lanes 5-7 from Grp15; and lanes 8-10 from Grp37. Lanes 1,3 and 4 have similar banding pattern; lane 2 has totally different banding pattern not matching with any other lane; lanes 5-7 have similar banding pattern but lane 7 (green arrow) seems to have very high concentration of the protein, giving it a smear look; lanes 8-10 seem to have similar banding pattern. This suggests that accessions within a group tend to have a similar banding pattern, which corroborates with the accession grouping with allele matching.

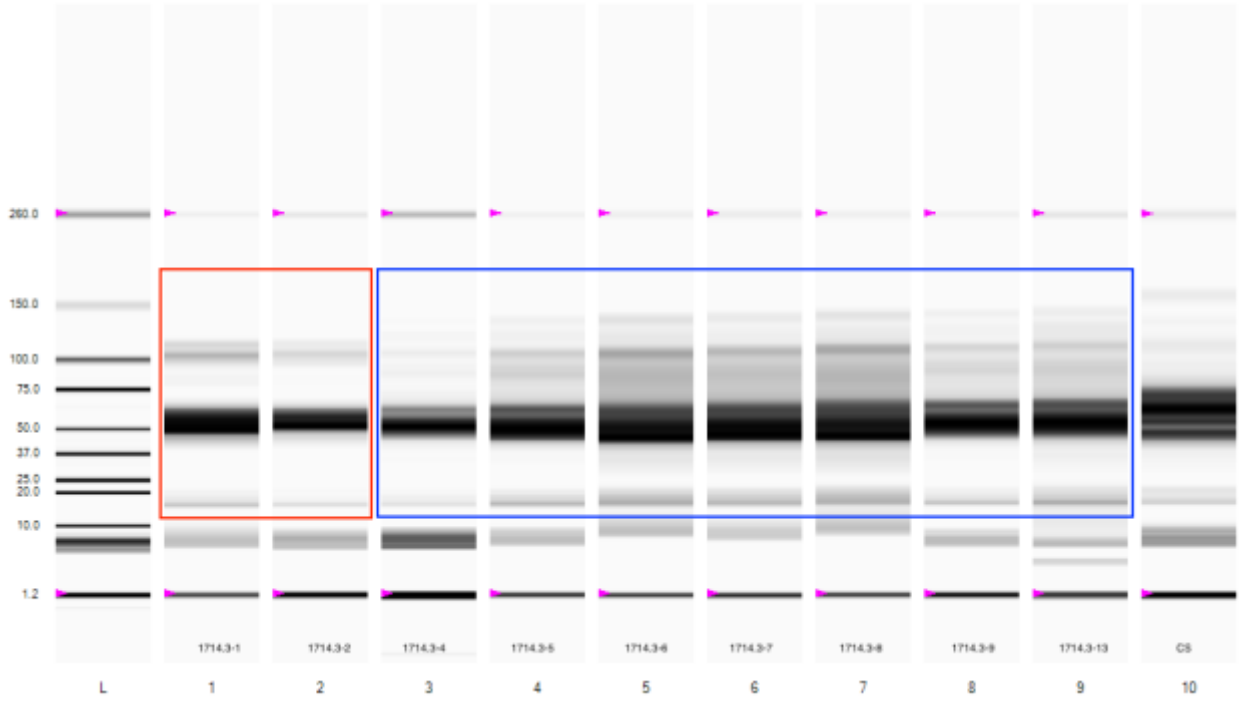

Figure 2.7. Virtual gel image of gliadin profiling for heterogeneous accession TA1714.

First two lanes (red box) have a similar banding pattern forming a group, and lanes 3-9 (blue box) have similar banding pattern with minor differences. Lane 10 is Chinese spring wheat for control. The different patterns between red and blue box samples presents an evidence that the samples came from a heterogeneous seed source.

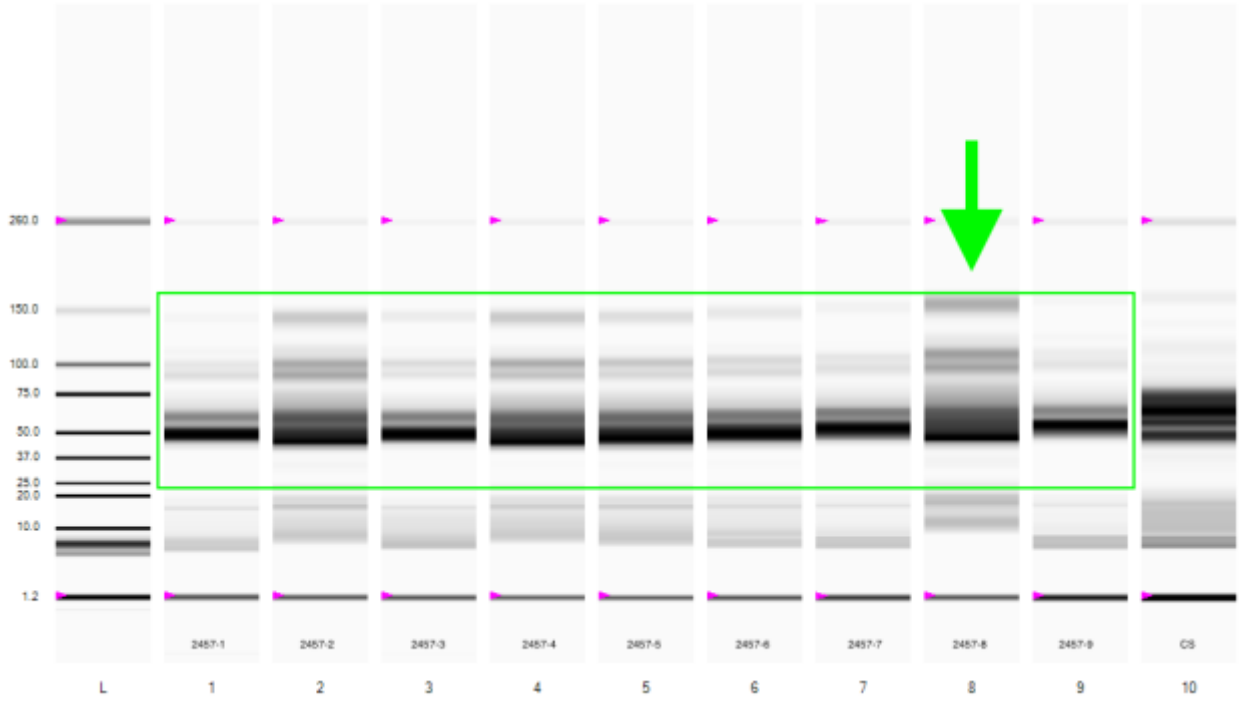

Figure 2.8. Virtual gel image of gliadin profiling for homogeneous accession TA2457.

With minor differences, banding pattern for lanes 1-9 (green box) look similar. Sample in lane 8 (green arrow) does appear to have a similar banding pattern but possibly has higher extracted protein concentration that gives it a smeared look.

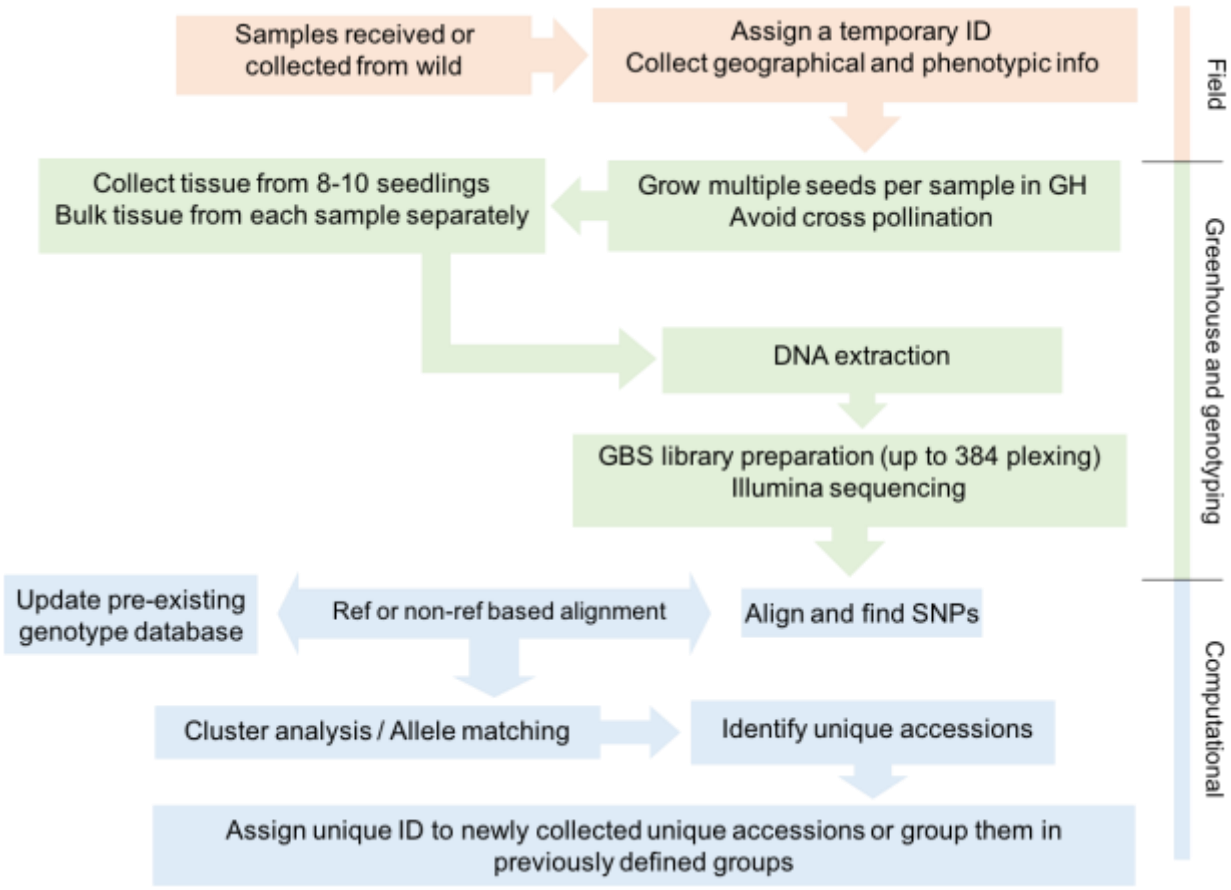

Figure 2.9. Future germplasm collection and management strategy to avoid the accumulation of redundant germplasm accessions.

Chapter 3 - Assessing genomic diversity in *Aegilops tauschii*

Abstract

Aegilops tauschii, the D-genome donor of bread wheat, *Triticum aestivum,* is a storehouse of genetic diversity and is an important resource for future wheat improvement. *Ae. tauschii* is split into two sub-lineages, L1 and L2, the latter being the wheat D-genome donor. A great amount of genetic diversity remains untapped in *Ae. tauschii*. Genomic and population analysis of 553 *Ae. tauschii* and 145 wheat accessions was performed by using 23,394 high quality SNPs. STRUCTURE, PCA and cluster analysis showed strong population differentiation of L1 and L2 within *Ae. tauschii*. L2 accessions had greater allelic diversity of the two sub-populations and wheat accessions had the least allelic diversity. Both sub-populations also showed differentiation with L1 being driven by longitudinal gradient and L2 differentiated by elevation. There has previously been little reported on natural hybridization between L1 and L2. We found nine putative natural inter-lineage hybrids in this study as admixture in the STRUCTURE analysis and containing many lineage-specific private SNP alleles from both lineages. One natural hybrid was confirmed as a recombinant inbred between the lineages. To facilitate the use of *Ae. tauschii* in wheat improvement, a MiniCore of 40 accessions has been developed based on genotypic, phenotypic and geographical data. MiniCore captures 95% allelic diversity from the collection.

Introduction

Aegilops tauschii, also known as goat grass, is the diploid donor of the D-genome of hexaploid bread wheat (*Triticum aestivum* L.). *Ae. tauschii* is native throughout the Caspian Sea region and into central Asia and China. Natural hybridization of tetraploid wheat and *Ae. tauschii* about 8,000-10,000 (Renfrew, 1973; reviewed in Bell, 1987) years ago led to the formation of

hexaploid wheat with *Ae. tauschii* contributing many genes that expanded the climatic adaption and improved bread making quality (Kihara 1944; McFadden & Sears 1946; Yamashita *et al.*, 1957; Kerber & Tipples, 1969; Lagudah *et al.*, 1991). However, during bread wheat evolution, only a handful of *Ae. tauschii* accessions from a narrow region hybridized with wheat leading to a narrow genetic base of the wheat D genome. Multiple studies have corroborated this, showing that the D-genome of wheat has the least genetic diversity as compared to its counterparts, A and B genomes (Kam-Morgan *et al.*, 1989; Lubbers *et al.* 1991; Akhunov *et al.*, 2010). However, huge amount of genetic diversity is present in this wild donor of the D-genome (Naghavi *et al.*, 2009).

With a pressing need to develop better yielding wheat varieties to feed a growing population and adapt to a changing climate, *Ae. tauschii* is a valuable source of novel alleles for wheat improvement (Kihara, 1944; Lagudah *et al.*, 1991). *Ae. tauschii* harbors considerable genetic diversity relative to the wheat D-genome and is split into two subspecies known as *Ae. tauschii* ssp. *tauschii* (Lineage 1; L1) and ssp. *strangulata* (Lineage 2; L2). The L2 ssp. *strangulata* is known to be the D-genome donor (Jaaska, 1978; Nakai, 1979; Nishikawa *et al.*, 1980; Jaaska, 1981). Ssp. *tauschii* is further split into three varieties- *typica*, *anathera* and *meyeri*, whereas ssp. *strangulata* is monotypic. Phenotypic classification of these subspecies and especially to varieties is challenging, therefore phenotypic data often poorly correlate with genetic classification (Lubbers *et al.* 1991; Dvorak *et al.*, 1998).

The narrow genetic base of modern bread wheat is a looming detriment to future wheat improvement and development of improved, climate resilient cultivars. Genetic diversity present in *Ae. tauschii* has been utilized via synthetic hybridization of tetraploid wheat and wild *Ae. tauschii*, but there is a still huge amount of untapped genetic diversity present in this species. In

this study, we characterized the full *Ae. tauschii* collection held at Wheat Genetics Resource Center at Kansas State University in Manhattan, KS, USA with the main objectives to: (i) genetically characterize the *Ae. tauschii* collection, (ii) study the population structure within *Ae. tauschii*, and (iii) develop a genetically diverse MiniCore set to facilitate the use of *Ae. tauschii* for wheat improvement.

Materials and methods

Plant material

This study included 574 (568 from Chapter 2 plus six new) *Ae. tauschii* accessions from Wheat Genetics Resource Center (WGRC) at Kansas State University (K-State) in Manhattan, KS, USA. Most of the *Ae. tauschii* accessions were collected in 1950s and '60s from 15 different countries by several explorers. In addition, a recent exploration was carried out by WGRC scientists in 2012 to fill the geographical gaps in the collection and sample more genetic diversity (Supplementary Fig. C.1 and Table C.1). Passport data, including longitude and latitude of the collection site, were available for most of the accessions and were plotted on the map to visualize the distribution (Fig. 3.1). To study the relationship between *Ae. tauschii* and hexaploid wheat (*Triticum aestivum* L.), 159 wheat varieties from a U.S. association mapping panel were also included in the study (Grogan *et al.*, 2016).

Plant tissue collection and genotyping-by-sequencing

A single plant for each accession were grown in 2"x2" pots in the greenhouse. Two to three inches of leaf tissue from single 2-3 weeks old seedlings were collected in 96-well tissue collection box, and stored at -80˚C until DNA extraction. Tissues were lyophilized in the lab for 24-36 hours, followed by genomic DNA extraction using Qiagen BioSprint 96 DNA Plant Kit (QIAGEN, Hilden, Germany). Extracted DNA was quantified with Quant-iT™ PicoGreen®

dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA). One random well per plate was left blank for quality control and library integrity. DNA samples were genotyping using genotyping-by-sequencing (GBS) (Poland, JA *et al.*, 2012). GBS libraries were prepared in 96 plexing using two restriction enzymes—a rare cutter *PstI* (5′-CTGCAG-3′), and a frequent cutter *MspI* (5'-CCGG-3') with a common reverse adapter ligated. Full protocol is available at the KSU Wheat Genetics website (http://wheatgenetics.org/download/send/3-protocols/74-gbs-protocol). GBS libraries were sequenced on 10 lanes on Illumina HiSeq2000 (Illumina, San Diego, CA, USA) platform at University of Missouri (UMC; Columbia, Missouri) and McGill Univesity-Génome Quebec Innovation Centre (Montreal, Canada) facility.

SNP genotyping and data filtering

Single nucleotide polymorphisms (SNPs) discovery and genotyping was performed in single step using Tassel 5 GBSv2 pipeline (Glaubitz *et al.*, 2014). The D-genome chromosomes from International Wheat Genome Sequencing Consortium (IWGSC) Reference Sequence v1.0 (RefSeq v1.0) assembly were used as the reference. Tassel was run with default settings and bowtie2 aligner in Linux HPC environment using a shell script. Genotypic data were processed in R statistical programming language (R Core Team, 2015) using custom R scripts. Population level SNP filtering was performed to remove multi-allelic SNPs. Remaining SNPs were filtered on per SNP and per individual basis. Minor allele frequency (MAF), per SNP and per individual proportion of missing data were computed. SNP markers with MAF less than 0.01 and missing data more than 20% were removed from the analysis. Further, SNPs with heterozygosity greater than 5% were removed because *Ae. tauschii* accessions are highly inbred. Finally, Fisher's exact test was run, with Bonferroni correction at family wise alpha level of 0.001, on all the SNPs to determine if the putative SNPs were from allelic tags as in Poland, J *et al.* (2012). Individual

samples with more than 20% missing SNP calls and more than 5% heterozygosity were also removed. Retained markers were used for further analyses.

Population structure analysis

Population structure analysis was performed with STRUCTURE software (Pritchard *et al.*, 2000) and Principal component analysis (PCA). Initial STRUCTURE was run with all filtered SNPs at K = 2 to partition L1 and L2 accessions. Per the developer recommendation for computational efficiency, final STRUCTURE analysis was performed using 10,000 randomly selected SNPs for K ranging from two to eight with three iterations of each K using admixture model with default settings except 10,000 burnins and 10,000 Monte Carlo Markov Chain iterations. To ensure the label collinearity for multiple iterations of each K run, STRUCTURE results were processed using *CLUMPAK* package (Jakobsson & Rosenberg, 2007; Kopelman *et al.*, 2015) and plotted using *Distruct* program (Rosenberg, 2004). Optimal K value was obtained using delta K method (Evanno *et al.*, 2005) using STRUCTURE HARVESTER (Earl, 2012).

Phylogenetic cluster analysis was performed in R language. Genetic distances were computed using 'dist' function with Euclidean method. Distance matrix was converted to a phylo object using 'ape' package (Paradis *et al.*, 2004). Using 'phyclust' package (Chen, 2011), a neighbor joining unrooted tree was plotted to indicate subpopulation clusters and identify tentative cryptic outliers that were not identified phenotypically. Cluster analysis was performed using default parameters in 'dist', 'ape' and 'phyclust'.

PCA was performed in R language. Eigenvalues and eigenvectors were computed using 'A' matrix output of 'rrBLUP' (Endelman, 2011). First three eigenvectors were plotted as three principal components to observe clustering. All analyses were performed separately for *Ae. tauschii* only to detect subpopulation, and with wheat to study the wheat-*Ae. tauschii*

relationship. PCA L1 and L2 sample coloring was done based on the STRUCTURE partitioning of two lineages. To find the best variables explaining the differentiation within lineages, samples were colored based on all the variables one at a time. Because of the continuous nature of the variables, a gradient coloring scheme was adopted to remove any bias due to arbitrary cut off boundary that is usually practiced for categorical variables.

Genetic diversity analysis

Nei's diversity index is the measure of average heterozygosity over multiple SNPs in a given population (Nei, 1973). This index was computed for the whole population, and separately for L1, L2, wheat, and combined for L1 and L2. Additionally, pairwise F_{ST} between subpopulations, and minor allele frequency (MAF), were computed using custom R scripts. Pairwise F_{ST} were computed among L1, L2, and wheat in all combinations. MAF plots were plotted separately for L1 and L2.

L1-L2 hybrids

To confirm the admixture of nine accessions that show up as intermediate hybrids of L1 and L2 in the STRUCTURE analysis, we evaluated the genotypes of these putative hybrids for lineage specific private alleles. For each lineage, SNPs were identified with private alleles (MAF=0 in the other lineage). The lineage specific private alleles were identified and the private allele contribution of each hybrid determined. Allele matching was performed as described in Chapter 2 to find the closest related accession (putative parent) from L1 and L2 for each hybrid.

Genetically diverse representative core-set selection

A random set of 15,000 SNPs were used to select a representative core-set from the *Ae. tauschii* collection. This reduced set of SNPs was used because of computational efficiency of the software. The core-set was selected in two steps. For the first step, software package

PowerCore was used (Kim *et al.*, 2007). PowerCore selects the samples to retain most diverse alleles and removing redundant alleles by implementing advanced M (maximization) strategy, and was run on default settings. However, the number of accessions selected by PowerCore were still too large to be utilized in a public research program. Therefore, the number of selected accessions was further reduced by phenotypically guided selection using the available phenotypic data for Leaf rust composite, Stem rust race TTKSK (Rouse *et al.*, 2011) and Hessian fly biotype L resistance. The diversity captured by the MiniCore was assessed by the percent segregating SNPs (MAF > 0.05) present in the selected accessions as compared to the whole collection.

Results

Geographical distribution of *Ae. tauschii*

Ae. tauschii is mainly found around the Caspian Sea and in central Asia, but is found as far West as Turkey (Lon: 26.327362, Lat: 40.009735) and as far East as eastern China (Lon: 111.048058, Lat: 34.059486). Geographical origin data was known for most of the accessions (Fig. 3.1). The majority of the accessions come from Afghanistan, Iran and Azerbaijan (Fig. C.2). Kyrgyzstan, Syria, Pakistan, China and Russia are among the least represented countries in our collection. L1 is spread across *Ae. tauschii* geographical range, whereas L2 is only present around the Caspian Sea region (Fig. 3.1).

Genomic profiling

Genotyping-by-sequencing (GBS) generated 742,028 putative single nucleotide polymorphisms (SNPs) from a total of 733 samples consisting of 574 *Ae. tauschii* and 159 wheat lines. Removing multi-allelic SNPs reduced the number to 710,531. Per SNP missing data ranged from 1.6-97%. SNPs with less than 20% missing data were retained resulting in 115,004

SNPs. Further filtering was performed by removing SNP markers with minor allele frequency less than 0.01 to remove extremely rare alleles. Since these wild accessions and wheat lines are inbred, SNPs with heterozygosity greater than 5% were removed, which ranged from 0-96%. Finally, SNPs were filtered using Fisher's exact test to confirm alternate SNP tags were allelic. In addition, sample filtering was also performed to remove poor samples based on the amount of missing data and heterozygosity. Per sample missing data ranged from 0-100%, and heterozygosity ranged from 0-40% for *Ae. tauschii* and 0-2% for wheat lines. Sixteen individual samples (two *Ae. tauschii* and 14 wheat) with more than 80% missing SNP calls, and 19 additional samples (all *Ae. tauschii*) with more than 5% heterozygous SNP calls were also removed. This resulted in a dataset of 24,713 SNPs for 698 samples consisting of 553 *Ae. tauschii* and 145 wheat samples. Finally, removing 1,319 SNPs that were private to D sub-genome of wheat, 23,394 high quality SNPs were retained and used for further analyses.

STRUCTURE analysis

Randomly selected 10,000 SNPs were used to infer the ancestry of all samples. Bayesian clustering software STRUCTURE was used to run model based clustering. Global analysis was run for *Ae. tauschii* and wheat together for K ranging from two to eight with three iterations for each K (Fig. 3.2). Samples were pre-assigned labels based on their geographical origin, and this information was used for plotting the STRUCTURE analysis. At K=2, L1 and L2 split from each other within *Ae. tauschii* and wheat D-genome remained clustered with L2 of *Ae. tauschii*. Nine accessions showed a very distinct structural differentiation as admixture of L1 and L2 (Fig. 3.2). These nine accessions are hypothesized as the possible hybrids between L1 and L2 and were analyzed separately. Using STRUCTURE default "Estimated Ln Prob of Data", K=3 was determined to be the optimal K, where L1, L2 and wheat D-genome were completely separated

(Fig. 3.2). At K=4, however, L1 showed population differentiation where accessions from Armenia, Azerbaijan, Georgia, Syria, and Turkey shows a different pattern than the accessions from Afghanistan, China, Kyrgyzstan, Pakistan, Tajikistan, and Uzbekistan. Accessions from Iran showed mixture of accessions from these two groups. At K=5, L2 showed some differentiation where accessions from Iran differentiated from Armenia, Azerbaijan and Georgia. For K > 5 no further information was provided by the STRUCTURE analysis in terms of population differentiation. L1 showed no further differentiation and L2 shows accessions from Armenia, Azerbaijan Georgia, Syria, and Turkey with similar ancestry and accessions from Iran as a complete admixture. Therefore, we determined K=5 to be a secondarily optimal stratification level after the optimal K=3. After K=3 and being differentiated from L2, wheat did not show any population differentiation in the entire global analysis.

STRUCTURE analysis was also run only on *Ae. tauschii* to see if removing wheat would change any pattern of grouping (Fig. C.3). Delta K method chose the optimal K=2, where L1 and L2 differentiated strongly. The same group of nine accessions as possible hybrid was evidenced as admixture of L1 and L2. At K=3, L1 showed the same population differentiation. Accessions from the eastern side of Caspian Sea differentiated from the western side. At K=4, L2 Iranian accessions showed admixture and differentiate from other accessions. At K=5, accessions from Turkmenistan in L1 started to show some admixture. At K=6, Pakistani accessions completely separates out from other L1 accessions. L2 did not show any differentiation after K=2. In contrast, L2 showed a weak population differentiation in presence of L1 with accessions from Iran showing a different differentiation pattern from the rest of the accessions (Fig. C.3). At K=4, the Iranian accessions started to show admixture. At K > 4 no more useful information was provided by the analysis for L2 differentiation.

In contrast to the *Ae. tauschii* accessions, wheat D-genome did not show any clear signals for population differentiation (Fig. 3.2). Some samples did show an admixture profile starting at K=2 and above. Starting at K=4, all samples started to show admixture and finally at K=6, all samples show admixture pattern, hence no significant population differentiation was detected in our wheat samples.

Principal component and cluster analysis

PCA was run as a second approach to cluster accessions and detect subpopulations. The same set of 23,394 *Ae. tauschii* specific SNPs were used for PCA. The defined lineages for *Ae. tauschii* individuals identified by STRUCTURE analysis was used to color the accessions in PCA (Supplementary Fig. C.4 and C.5) and phylogenetic cluster analysis (Fig. 3.3). Principal component analysis was performed separately for two datasets- *Ae. tauschii* with wheat, and *Ae. tauschii* only. As expected, the population differentiation observed in STRUCTURE was confirmed as three distinct groups of L1, L2 and wheat were observed in the first two components of the PCA (Fig. C.4). PCA1 explained 54% of the variation separating L1 and L2. PCA2 explained 9% of the variation and separates out wheat from L2 of *Ae. tauschii*. Corroborating previous reports, the wheat D-genome was again observed to be more closely related to L2 accessions.

PCA with only the *Ae. tauschii* accessions, again confirmed the strong population differentiation between two *Ae. tauschii* lineages, L1 and L2. In this analysis, PCA1 explained 52% of the variation in the dataset (Fig. 3.4 and C.5). When analyzed in the absence of wheat, L1 shows a strong within lineage differentiation on the second principal component explaining 4% of the variation, and L2 on the third principal component explaining 4% of the variation. Within lineages, the samples were colored based on all the variables individually. L1

differentiation was correlated to the longitudinal geographical origin of accessions with an east-west gradient relative to the Caspian Sea. After removing seven outlier accessions, when the longitudes of L1 accessions are plotted against PC2, it clearly shows the upward trend with PC2 separating eastern and western accessions (Fig. 3.5). On the third principal component, population differentiation was also observed, which corresponded to the altitude of origin of the L2 accessions in reference to the sea level. PC3 vs. altitude plot also shows a clear trend with PC3 separating the accessions according to their altitude, however there are few outliers present on the both ends (Fig. C.6). Generally, lower elevation accessions clustered together separately from the higher elevation accessions. We found that the strongest differentiation between L2 clusters was at around 150m above sea level. Overall the PCA results were in strong agreement with the population differentiation observed in STRUCTURE.

As a final assessment of population structure, Cluster analysis was performed by computing genetic distances among accession using Euclidean method. An unrooted tree in this cluster analysis splits samples into three distinct clades- L1, L2 and wheat (Fig. 3.3). Wheat and L2 are more closely related than wheat and L1, and L1 and L2. L1 and L2 further shows two clades within that could again be attributed to longitudinal variation from the Caspian Sea and elevation, respectively. Wheat essentially did not show any differentiation within.

Admixed *Ae. tauschii* accessions are L1-L2 hybrids

Nine accessions showed up in STRUCTURE, PCA and cluster analysis as admixture of *Ae. tauschii* L1 and L2. To test their origin as hybrids between L1 and L2 accessions, private alleles in both lineages were filtered and tested in the hybrid samples. Count for lineage specific alleles contributed by L1 ranged from 843 to 1684, and by L2 ranged from 589 to 1236. On average, L1 contributed 974 alleles, and L2 contributed 1095 alleles genome-wide across all

hybrids. Out of nine putative hybrid samples, TA3429 was confirmed as a bi-parental recombinant inbred line between L1 and L2 accessions, in which the chromosomal segments from L1 and L2 were clearly demarcated without overlapping (Fig. 3.6). However, other samples showed no clear patterns of genome contributions (Supplementary Fig. C.7), contrary to expected blocks of lineage specific origin.

To find potential L1 and L2 parents of each putative hybrid, allele matching was performed. SNPs with lineage specific private alleles were used to find the closest accession from each lineage. Lowest and highest percent identity was found to be 59.19% and 76.34%, respectively, between a pair of hybrid and L1 accessions. Similarly, the lowest and highest percent identity between any pair of hybrid and L2 accessions was found to be 56.59% and 62.65%, respectively. List of putative hybrids with highest matching accessions is summarized in Table 2.

Genetic diversity

Nei's diversity index was computed using all SNPs separately for *Ae. tauschii* L1, L2, possible hybrids, wheat and *Ae. tauschii* collection combined. Highest Nei's diversity index was observed for L2 = 0.1602 followed by L1 = 0.1112, and wheat of 0.0347. Higher values of the Nei's index indicates greater allelic diversity in a given population. Combined Nei's index for *Ae. tauschii* was 0.3083 and the whole dataset, including wheat, was 0.3539.

To evaluate population differentiation between the different pairs of population groups, pairwise F_{ST} statistics were computed between L1, L2, L1-L2 hybrids and wheat. Highest F_{ST} were observed between L1 and wheat, followed by wheat and L2, and wheat and L1-L2 hybrids (Table 1). The population differentiation between L1 and wheat also supports the large number of novel of alleles found in this lineage that are absent from the wheat pool.

MAF was computed and plotted separately for L1, L2 and jointly for both lineages (Supplementary Fig. C.8). Individually, MAF spectrum for L1 and L2 showed an expected distribution with majority of alleles present at very low frequency (panel A and B, Fig. C.8). Joint distribution of L1 and L2 MAF revealed that majority of the alleles segregating in one lineage were close to fixation in the other lineage (panel C, Fig. C.8).

Core-set selection

Genetically diverse core-set was selected using software package PowerCore that implements advanced M (maximization) strategy to select diverse accession by reducing allelic redundancy and keeping the allele frequency spectrum similar. Initially 107 *Ae. tauschii* accessions were selected using advanced M strategy implemented in PowerCore (Table C.1). These accessions were then plotted on a phylogenetic tree and reduced using known phenotypic information on disease and insect resistance to get the size of this core to a manageable number. This selection was guided by phenotypic data for resistance to Leaf rust composite, Stem rust TTKSK race and Hessian fly biotype L. Other factors, such as the available geographical origin and the history of their previous use in genetic mapping, were also taken into account to pick the representative accessions. Finally, 40 accessions were selected to comprise a MiniCore that is distributed uniformly across the WGRC *Ae. tauschii* collection (Fig. C.9). Nei's diversity index computed for the MiniCore (0.2933) compared to the whole collection (0.3083) suggests the presence of allelic richness in the MiniCore. Also in the MiniCore, we were able to retain 16,408 segregating SNPs (MAF > 0.05) out of 17,274 from the whole *Ae. tauschii* collection. By reducing the collection size by over 10 fold, we were still able to capture 94.9% of the segregating alleles present in the whole WGRC collection. MiniCore consists of 29 accessions from L1 and 11 accessions from L2 of *Ae. tauschii*.

Discussion

Ae. tauschii distribution

Caspian Sea region is thought to be the center of origin of *Ae. tauschii*. Most of the accessions in our collection were also sampled from this region (Fig. 3.1). L2 of *Ae. tauschii* is spread on a narrow longitudinal range from northeastern Syria to northeastern Iran spanning a distance of 1625 km, whereas L1 is found from southern Turkey to northwestern China, spanning over 4000 km. Most of the accessions came from other genebanks, however, to fill up the geographical gaps, a recent exploration was conducted in 2012 by WGRC researchers (blue dots, Fig. C.1). Both lineages are found overlapping at similar altitudes, with generally L1 accessions found at higher altitudes than L2 (Fig. C.10). Majority of L1 and L2 accessions fall in the similar latitude distribution, but some L1 accessions were widely spread.

SNP discovery and ascertainment bias

Using D-genome specific chromosomes from IWGSC RefSeq v1.0 as a reference, GBS produced 18,127 high quality markers to assess genetic diversity in the collection. As we did not use any prior SNP information to call SNPs, we expect the ascertainment bias be minimal. However, we used Chinese Spring wheat as a reference genome to call SNPs, therefore, we do expect some loci missing in our dataset because of potentially missing *Ae. tauschii* chromosomal segments in the wheat reference genome. Because the goal of this project was not to assess any specific genomic region, using wheat as a reference genome should not pose a problem as most of the sequence order is generally conserved among close relatives (International Wheat Genome Sequencing, 2014). Wheat is a hexaploid species, and the presence of A and B genomes could result in false SNPs on the reference D-genome. However, the low level of heterozygosity (less than 3%) in all the wheat lines presents evidence that only D-genome sequence reads were

mapped uniquely on the reference and we did not inflate SNP calling by offsite mapping of wheat sequence tags from homoeologs.

Population structure analysis

Global population structure analysis was performed using default parameters in STRUCTURE program, which normally assumes that marker data be in Hardy–Weinberg equilibrium and not in linkage disequilibrium. However, starting version 2, authors noted that the program was able to handle mild departures from these assumptions. The default admixture model is flexible to accommodate complexities of the real data and deal with hybrid zones. Therefore, SNP filtering was not performed to remove markers that violate those assumptions. Global analysis revealed K=3 to be the optimal level of population differentiation. This is expected as *Ae. tauschii* has been reported to have a strong population structure into two groups with the wheat D-genome forming a third group. At K=2 the wheat D-genome grouped with L2 supporting that this subpopulation of *Ae. tauschii* was the D-genome donor of hexaploid wheat. A small group of nine accessions showed up as hybrids of L1 and L2. At K=4, *Ae. tauschii* L1 showed intra-lineage population differentiation in accordance with relative position of East or West of the Caspian Sea. This was also clear in the principal component analysis where L1 was differentiated by PCA2 along longitudinal gradients. Iranian accessions did not show clear population differentiation by falling into the eastern or western group but rather show admixture. Iran is at the center of origin for *Ae. tauschii* and could be seen as a transition region for the East and West clades of L1. The majority of the L2 accessions occur in Azerbaijan and Iran, both of which are on one side of the Caspian Sea, therefore longitudinal gradient did not explain the weak population structure within L2 at K=5. However, we found that this population differentiation could be attributed to the elevation of the origin of L2 accessions where

accessions originating at less than 150m above sea level cluster separately from the accessions from more than 300m above sea level (PCA Fig. 3.4 and C.6). The wheat lines in this study did not show any population differentiation (Fig. 3.2).

L1-L2 hybrids

Hybrids between L1 and L2 of *Ae. tauschii* are rare and have been the subject of limited reports. Wang (2013) found two accessions falling in between L1 and L2. Based on haplotype distribution similarity and close geographical proximity of origin, they concluded that these two accessions originated from the hybridization of a single L2 plant with an L1 plant. In the present study, we found nine accessions as admixture between L1 and L2 in STRUCTURE, PCA and cluster analyses. Using the SNPs with lineage specific alleles, the allele matching of these nine accessions did not yield 99% match or more identity with any single L1 or L2 accession, but matched identity at intermediate levels. This suggests that the parental accessions were not present in our collection. However, when we looked at the L1 and L2 accessions that had highest identities with putative hybrids, there were only three from each lineage. This could suggest that the natural hybridization of L1 and L2 accessions is uncommon, and these hybrids are possibly originated from one or few of these rare events. These lineages co-exist with each other therefore they are not isolated by distance, therefore, possibly they are reproductively isolated given their inbreeding nature. Similar pattern of reproductive isolation and rare hybridization was reported in rice landraces (Huang *et al*., 2010), and switchgrass (Grabowski *et al*., 2014). This is consistent with the findings of Wang *et al.* (2013), where they suggested a single hybridization event might have resulted in the two accessions of L1-L2 admixture individuals in their data. The distribution of L1 and L2 private alleles in these hybrids supports our hypothesis that these accessions are hybrid (Fig. 3.6 and Supplementary Fig. C.7). One accession, TA3429 shows a

pattern as expected in a recombinant inbred of a hybrid between two accessions from opposite lineages. These natural hybridization events could produce novel genetic variation by reshuffling and assembling the genome for a wide range of traits. However, the dearth of L1-L2 hybrids in nature and the strong genetic differentiation between L1 and L2 suggests the reproductive isolation of these two lineages. Similar conclusions have also been reported in other studies using different marker systems (Mizuno *et al.*, 2010; Sohail *et al.*, 2012), however, using GBS, we reported the confirmation of one L1-L2 hybrid for the first time.

Genetic diversity

Wheat had the lowest Nei's index, which is expected because of its domestication and polyploidization, compared to its wild progenitor, *Ae. tauschii*. Reduction in genetic diversity has also been reported in cotton as a result of change in ploidy level (Iqbal et al., 2001). Wheat lines in our study also represent a relatively narrow collection of US winter wheat, leading to the lowest Nei's index. Highest Nei's index was observed for L2, followed by L1. This can be attributed to the differences in distribution of L1 and L2 across their natural habitat. L1 is distributed across the longitudinal gradient, whereas L2 is distributed across the elevation gradient. Latitude is known to affect the temperature with cooler temperatures away from the equator (Rind, 1998), but the latitude distribution for L1 and L2 was similar for majority of accessions except a few outliers (panel C, Fig. C.10). Therefore, the expected effect of latitude affect should be minimal. Longitude distribution for L1 was more extensive as compared to L2 (panel B, Fig. C.10). As shown in Fig. 3.1, the majority of the L2 accessions are distributed around the Caspian Sea as compared to very few L1 accessions. Therefore, the longitude effect is more pronounced in L1 than L2. Moreover, the elevation distributions for L1 and L2 were also different (panel A, Fig. C.10), with more L2 accessions growing at lower elevation. Elevation is

known to have an effect on the temperature (Körner, 2007). Therefore, L2 accessions might have selected alleles to survive in different temperatures. Combined *Ae. tauschii* had higher Nei's index as compared to any single lineage, which is expected because the allelic diversity is assayed in the whole collection.

Genetically diverse representative MiniCore

Accessing the genetic diversity present in wild relatives can be a challenging task for breeders due to the large number of accessions and confounding physiology of the wild plants. Wild accessions with overall poor phenotype could be the source of agronomically important alleles. Efficient use of germplasm collections can often be facilitated through a targeted subset of the total accessions that is optimized to capture a maximum amount of the total diversity in a minimum number of accessions. To facilitate the use of *Ae. tauschii* accessions in wheat breeding, we selected only 40 accessions to develop a smaller MiniCore set that captures 95% of the segregating alleles from the whole collection. MiniCore was carefully selected from both the lineages of *Ae. tauschii* but the main focus was to target more from L1. This is because L1 is a reservoir of untapped genetic diversity that has not been leveraged by the breeders. L2 accessions were chosen because this lineage is the source for many of the diseases and insect resistance. These accessions can be utilized to bring in novel genetic variation for wheat rusts, insect resistance, heat and drought tolerance to produce climate resilient wheat varieties. This MiniCore consisting of genetically diverse accessions was selected with an objective to broaden the genetic base of wheat D-genome. However, in future, the selection can be optimized based on the recombination rate and the distribution of *Ae. tauschii* regions that are already introgressed in the wheat D-genome.

Acknowledgements

This study was conducted under the auspices of the Wheat Genetics Resource Center (WGRC) Industry/University Collaborative Research Center (I/UCRC) through support of industry partners and partially funded by NSF grant contract (IIP-1338897). We would also like to acknowledge the contribution of Industry Advisory Board to improve this manuscript with their valuable comments and suggestions.

References

Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, Coleman-Derr D, Conley EJ, Crossman CC, Deal KR, *et al.* 2010. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* **11**: 702.

Bell G 1987. The history of wheat cultivation. Wheat Breeding: *Springer*, 31-49.

Chen W-C. 2011. *Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm.* Iowa State University.

Dvorak J, Luo MC, Yang ZL, Zhang HB. 1998. The structure of the *Ae. tauschii* genepool and the evolution of hexaploid wheat. *Theoretical and Applied Genetics* **97**(4): 657-670.

Earl DA. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation genetics resources* **4**(2): 359-361.

Endelman JB. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome* **4**(3): 250-255.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**(2): e90346.

Grabowski PP, Morris GP, Casler MD, Borevitz JO. **2014**. Population genomic variation reveals roles of history, adaptation and ploidy in switchgrass. *Molecular Ecology* **23**: 4059–4073.

Grogan SM, Anderson J, Baenziger PS, Frels K, Guttieri MJ, Haley SD, Kim K-S, Liu S, McMaster GS, Newell M. 2016. Phenotypic plasticity of winter wheat heading date and grain yield across the US Great Plains. *Crop Science* **56**(5): 2223-2236.

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, *et al.* 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics* **42**: 961–967.

International Wheat Genome Sequencing C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**(6194): 1251788.

Iqbal MJ, Reddy OUK, El-Zik KM, Pepper a. E. **2001**. A genetic bottleneck in the 'evolution under domestication' of upland cotton Gossypium hirsutum L. examined using DNA fingerprinting. *TAG Theoretical and Applied Genetics* **103**: 547–554.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**(14): 1801-1806.

Jaaska V. 1978. NADP-dependent aromatic alcohol dehydrogenase in polyploid wheats and their diploid relatives. On the origin and phylogeny of polyploid wheats. *Theoretical and Applied Genetics* **53**(5): 209-217.

Jaaska V. 1981. Aspartate aminotransferase and alcohol dehydrogenase isoenzymes: Intraspecific differentiation in*Aegilops tauschii* and the origin of the D genome polyploids in the wheat group. *Plant Systematics and Evolution* **137**(4): 259-273.

Kam-Morgan L, Gill B, Muthukrishnan S. 1989. DNA restriction fragment length polymorphisms: a strategy for genetic mapping of D genome of wheat. *Genome* **32**(4): 724-732.

Kerber E, Tipples K. 1969. Effects of the D genome on milling and baking properties of wheat. *Canadian Journal of Plant Science* **49**(3): 255-263.

Kihara H. 1944. Discovery of the DD-analyser, one of the ancestors of *Triticum* vulgare. *Agric Hortic* **19**: 13-14.

Kihara H, Tanaka M. 1958. Morphological and physiological variation among *Aegilops* squarossa strains collected in Pakistan, Afghanistan and Iran. *Preslia* **30**: 241–251.

Kihara H, Yamashita H, Tanaka M. 1965. Morphological, physiological, genetical, and cytological studies in *Aegilops* and *Triticum* collected in Pakistan, Afghanistan, Iran. Results of the Kyoto University scientific expedition to the Korakoram and Hidukush in 1955. Volume 1. Kyoto, Japan: Kyoto University.

Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG, Park YJ. 2007. PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* **23**(16): 2155-2162.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular ecology resources* **15**(5): 1179-1191.

Körner C. 2007. The use of 'altitude'in ecological research. *Trends Ecol Evol* **22**(11): 569-574.

Lagudah ES, Appels R, Brown AHD, Mcneil D. 1991. The Molecular-Genetic Analysis of *Triticum*-Tauschii, the D-Genome Donor to Hexaploid Wheat. *Genome* **34**(3): 375-386.

Lubbers EL, Gill KS, Cox TS, Gill BS. 1991. Variation of Molecular Markers among Geographically Diverse Accessions of *Triticum tauschii*. *Genome* **34**(3): 354-361.

McFadden FE, Sears ER. 1946. The origin of *Triticum* spelta and its free-threshing hexaploid relatives. *J Hered* **37**: 81 107.

Mizuno N, Yamasaki M, Matsuoka Y, Kawahara T, Takumi S. 2010. Population structure of wild wheat D-genome progenitor *Ae. tauschii* Coss.: implications for intraspecific lineage diversification and evolution of common wheat. *Mol Ecol* **19**(5): 999-1013.

Naghavi M, Aghaei M, Taleei A, Omidi M, Mozafari J, Hassani M. 2009. Genetic diversity of the D-genome in T. *aestivum* and *Aegilops* species using SSR markers. *Genetic Resources and Crop Evolution* **56**(4): 499-506.

Nakai Y. 1979. Isozyme variations in *Aegilops* and *Triticum*, IV. The origin of the common wheats revealed from the study on esterase isozymes in synthesized hexaploid wheats. *The Japanese Journal of Genetics* **54**(3): 175-189.

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* **70**(12): 3321-3323.

Nishikawa K, Furuta Y, Wada T. 1980. Genetic studies on alphaamylase isozymes in wheat. III. Intraspecific variation in *Aegilops* squarrosa and birthplace of hexaploid wheat. *The Japanese Journal of Genetics* **55**(5): 325-336.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**(2): 289-290.

Poland J, Endelman J, Dawson J, Rutkoski J, Wu SY, Manes Y, Dreisigacker S, Crossa J, Sanchez-Villeda H, Sorrells M, *et al.* 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *Plant Genome* **5**(3): 103-113.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* **7**(2): e32253.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**(2): 945-959.

R Core Team. 2015. R: A language and environment for statistical computing: R Foundation for Statistical Computing.

Renfrew JM. 1973. Palaeoethnobotany. The prehistoric food plants of the Near East and Europe.

Rind D. 1998. Latitudinal temperature gradients and climate change. *Journal of Geophysical Research: Atmospheres* **103**(D6): 5943-5971.

Rosenberg NA. 2004. DISTRUCT: a program for the graphical display of population structure. *Molecular ecology resources* **4**(1): 137-138.

Rouse MN, Olson EL, Gill BS, Pumphrey MO, Jin Y. 2011. Stem Rust Resistance in *Ae. tauschii* Germplasm. *Crop Science* **51**(5): 2074-2078.

Sohail Q, Shehzad T, Kilian A, Eltayeb AE, Tanaka H, Tsujimoto H. 2012. Development of diversity array technology (DArT) markers for assessment of population structure and diversity in *Ae. tauschii*. *Breeding science* **62**(1): 38-45.

Wang J, Luo MC, Chen Z, You FM, Wei Y, Zheng Y, Dvorak J. 2013. *Ae. tauschii* single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *New Phytol* **198**(3): 925-937.

Yamashita K, Tanaka M, Koyama M. 1957. Studies on the flour quality in *Triticum* and *Aegilops*. *Seiken Ziho* **8**: 20-26.

Table 3.1. Pairwise FST coefficients among L1, L2, L1-L2 hybrids, and Wheat. Higher values represent stronger population differentiation.

	L1	*L2*	*Wheat*	*L1-L2*
L1	-	0.6550	0.6621	0.4744
L2	0.6550	-	0.4948	0.4837
Wheat	0.6621	0.4948	-	0.3965
L1-L2	0.4744	0.4837	0.3965	-

Table 3.2. Putative hybrids and their closest L1 and L2 accessions. Numbers in parenthesis represent percent identity of the hybrid with the respective accession.

Putative hybrid	*Geographical origin*	*L1 accession*	*L2 accession*
TA10104	Georgia	TA2482 (59.28%)	TA2530 (62.57%)
TA10103	Georgia	TA2482 (59.45%)	TA2530 (61.95%)
TA10113	Turkmenistan	TA10932 (66.28%)	TA2527 (56.59%)
TA2576	Georgia	TA2482 (59.55%)	TA2530 (62.63%)
TA3429	-	TA1595 (76.34%)	TA2377 (57.33%)
TA10929	Georgia	TA2482 (59.67%)	TA2530 (62.31%)
TA10928	Georgia	TA2482 (59.55%)	TA2530 (62.65%)
TA2582	Georgia	TA2482 (58.91%)	TA2530 (61.9%)
TA2580	Georgia	TA2482 (59.19%)	TA2530 (61.95%)

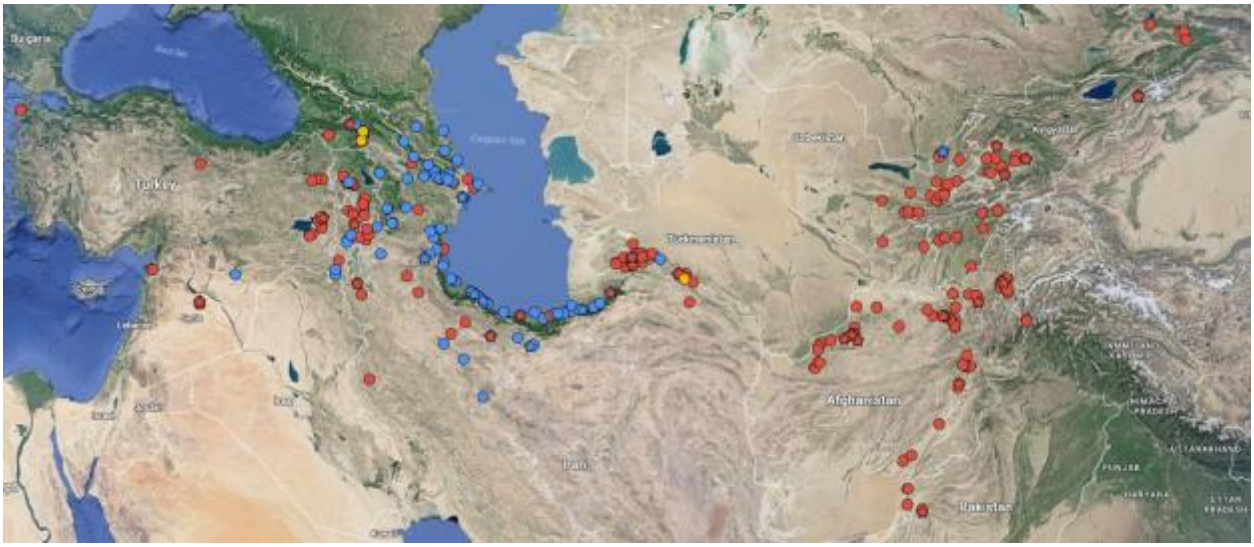

Figure 3.1. Geographical distribution of *Ae. tauschii* accessions.

Solid red circles represent Lineage 1 (L1), solid blue circles Lineage 2 (L2), red stars MiniCore from L1, blue stars MiniCore from L2, and solid gold circles putative hybrids.

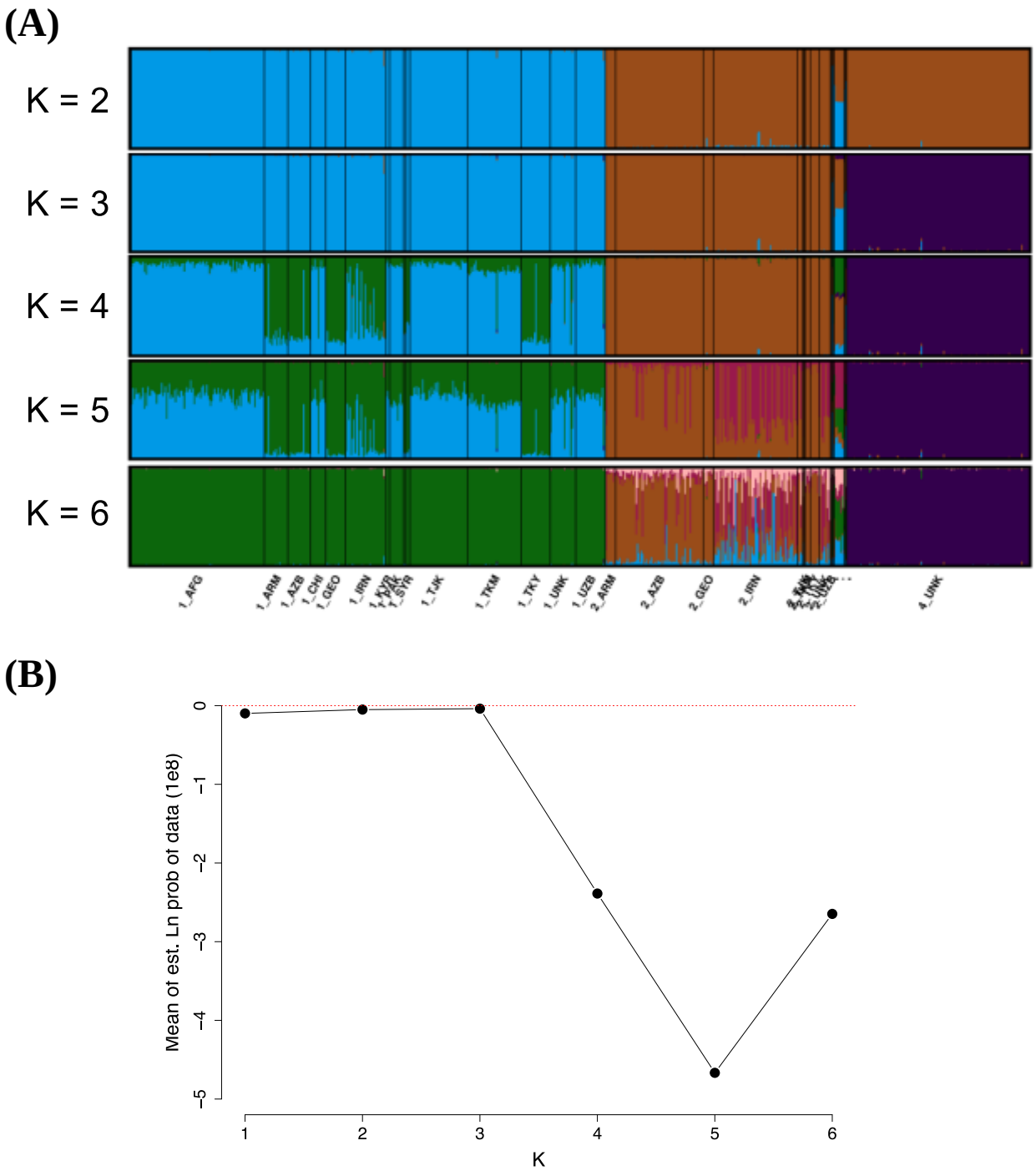

Figure 3.2. (A) Global STRUCTURE analysis for *Ae. tauschii* L1, L2, L1-L2 hybrids and wheat for K=1 to K=6. (B) Estimated Ln Prob of data showing the optimal K.

An additional color is added with each increase in the value of K. Each vertical bar represents an individual. A bar with only a single color represents its ancestry to a single population, and a mixture of colors represents admixture from different populations.

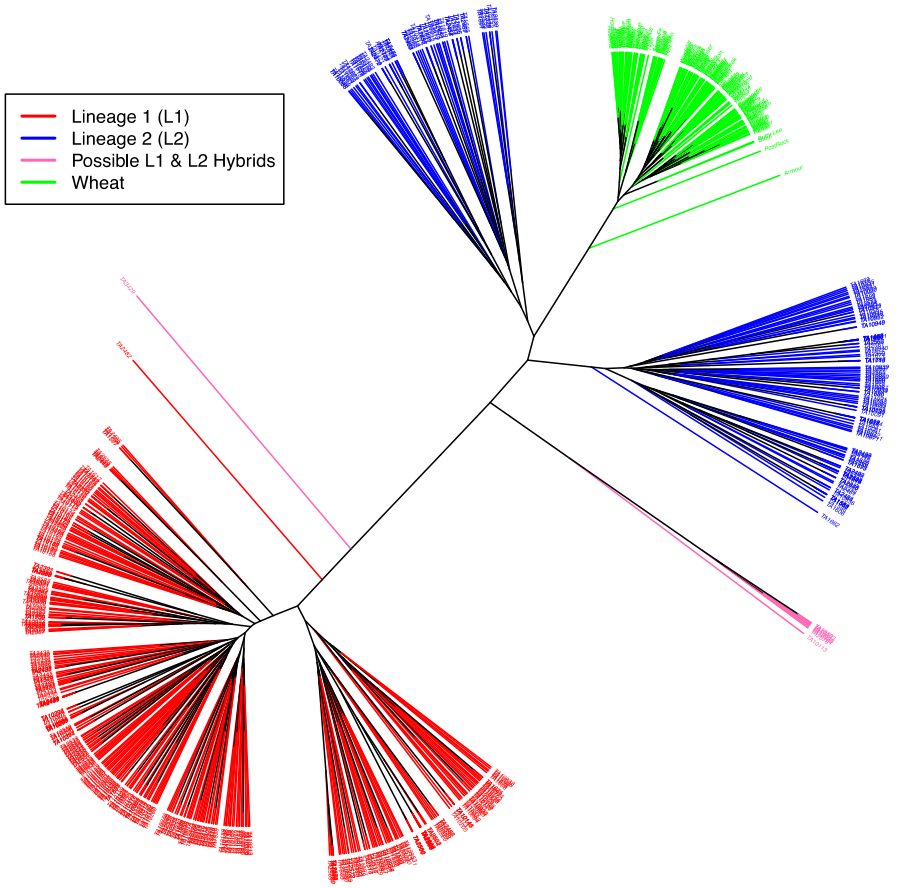

Figure 3.3. Neighbor-joining cluster analysis showing relationship between L1, L2, L1-L2 hybrids and wheat.

Red branches represent L1, blue L2 and green wheat. Wheat is closely related to L2 of *Ae. tauschii*. Putative hybrids cluster out separately and appear in between the two lineages.

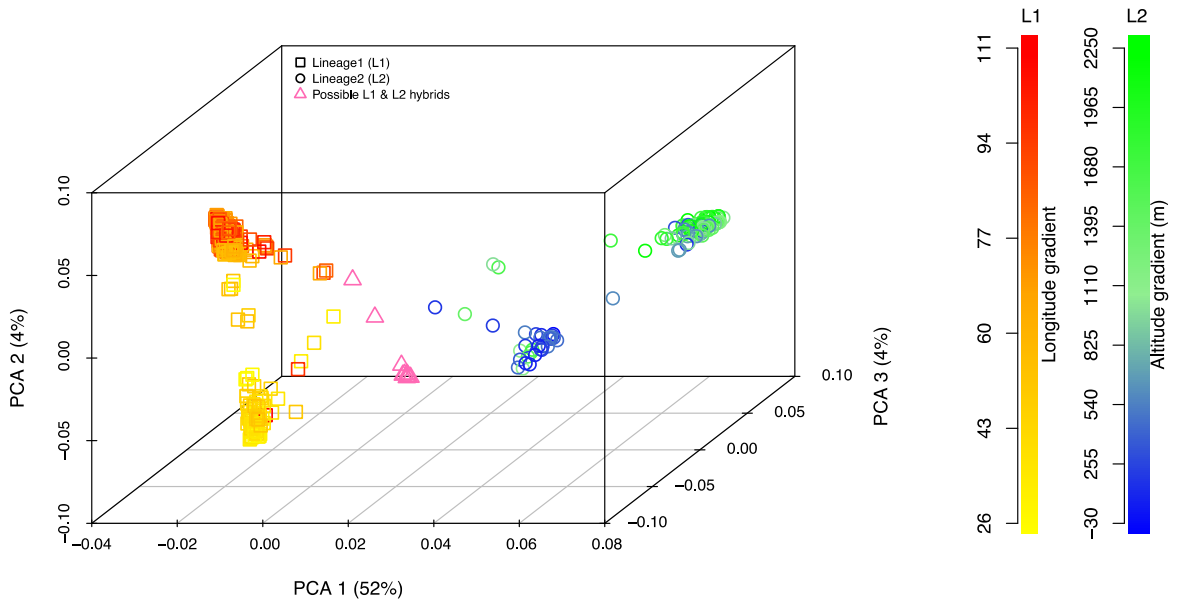

Figure 3.4. Three-dimensional principal component analysis plot for L1, L2, and putative L1-L2 hybrids.

Lineage1 (L1) is colored based on the longitudinal gradient and Lineage2 is colored with elevation gradient with reference to the sea level. Empty squares represent L1 (yellow-red gradient) and empty circles represent L2 (blue-green gradient). Hot-pink empty triangles represent putative L1-L2 hybrids. Legends for the color gradient are shown on the right side.

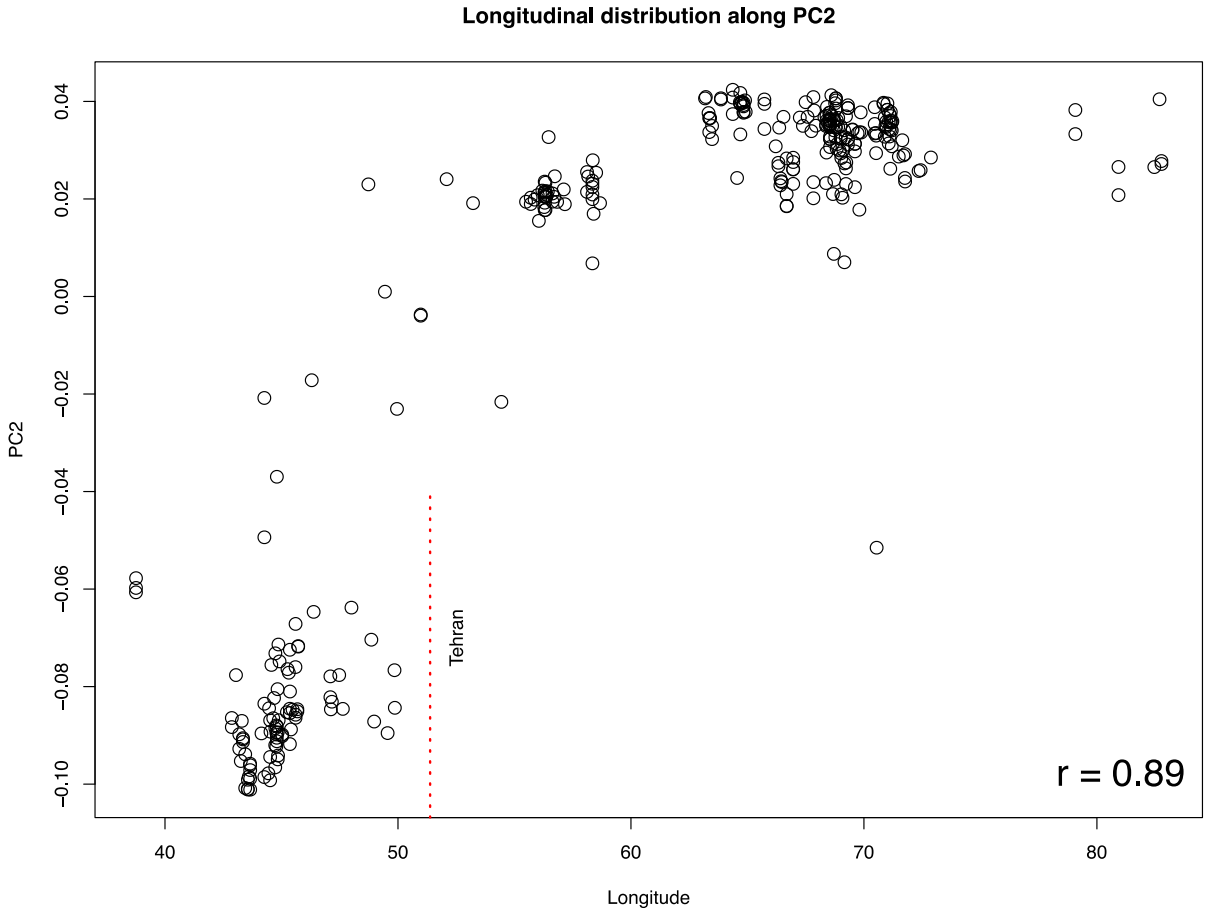

Figure 3.5. Scatterplot showing the relationship between L1 longitude and PC2.

Strong correlation between two variables is evident suggesting that PC2 is separating out western (lower longitudes) from eastern (higher longitudes) accessions. Correlation coefficient is shown at the bottom right corner. Vertical red dotted line marks the longitude of Tehran, Iran that demarcates the eastern vs. western accessions.

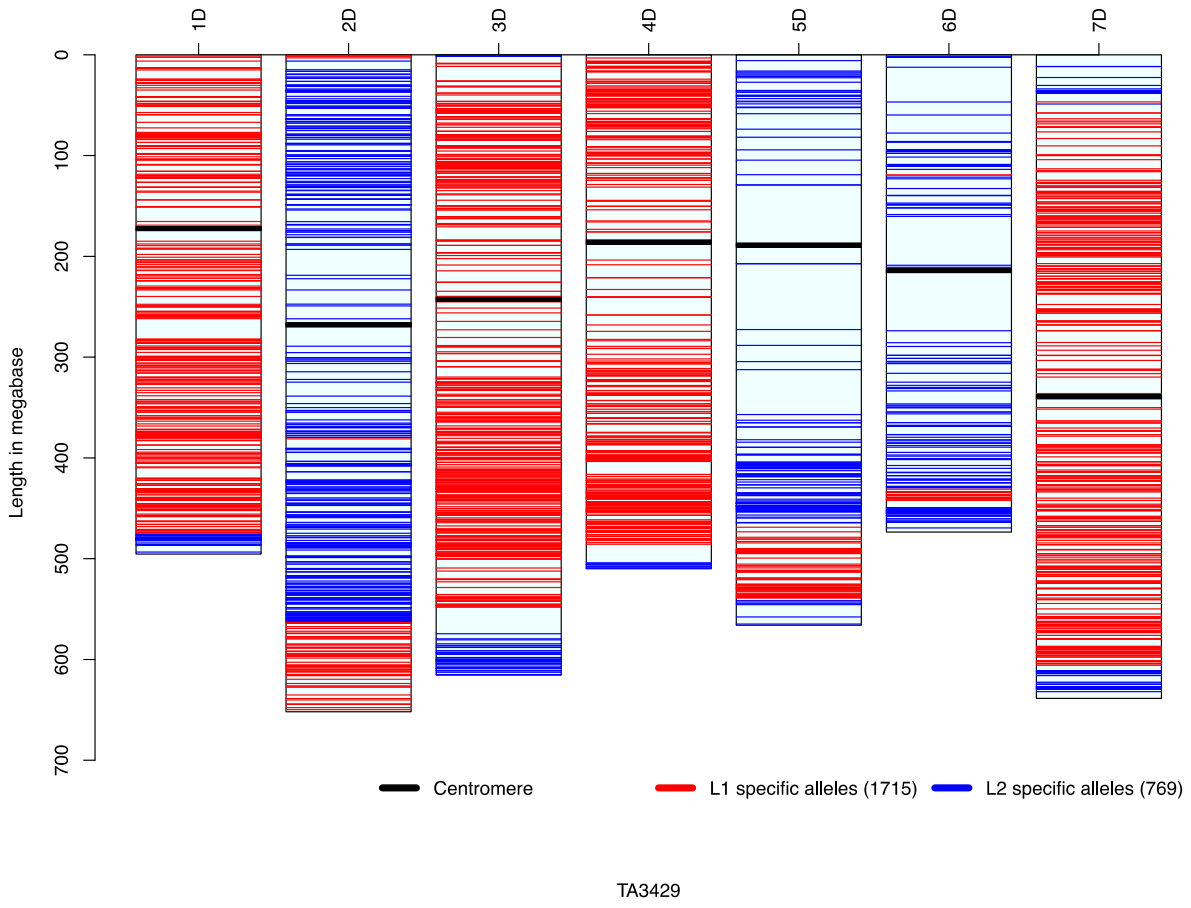

Figure 3.6. Distribution of lineage specific alleles for putative hybrid TA3429.

Red color represents L1 specific alleles, and blue represents L2 specific alleles. Numbers in parentheses in the legend represent the number of lineage specific alleles present in a given hybrid.

Chapter 4 - Genetic mapping of one new and three known resistance genes for Hessian fly in winter wheat

Abstract

Hessian fly (HF; *Mayetiola destructor* Say) is a devastating insect pest of wheat worldwide. About three dozen HF resistance genes have been identified, but lack of genetic mapping and linked molecular markers limit their utilization in wheat breeding and prevent molecular breeding approaches. Typical HF resistance genes follow the gene-for-gene model, that creates strong selection pressure causing rapid shifts in HF biotypes. Therefore, optimal breeding strategies require gene-pyramiding and deployment of multiple resistance genes, both of which require genetic positions and linked molecular markers. In this study, we mapped three previously known genes, *H5*, *H10* and *H13*, using genotyping-by-sequencing. In addition, we also identified and mapped a new introgression from *Ae. tauschii* in winter wheat cultivar 'Overley'. *H5* was mapped on the telomeric region of short arm of chromosome 6B. *H10* and *H13* were mapped at the same locus at the distal end of short arm of chromosome 6D. The gene introgressed from *Aegilops tauschii*, designated here as *H2147*, was mapped on the distal end of long arm of chromosome 3D. Utilizing R-gene enrichment sequencing (RenSeq), we identified a region with the introgression carrying seven candidate NBS-LRR genes for *H2147*. Knowing the positions of HF resistance genes with marker information will facilitate their use in insect resistance wheat breeding, and will allow breeders to stack multiple genes in one cultivar background, which will also reduce the strong selection pressure on the HF populations, avoiding resistance breakdown and improving durability of resistance.

Introduction

Hessian fly (HF; *Mayetiola destructor* Say; Diptera: Cecidomyiidae) is a serious hemibiotrophic pest of wheat (*Triticum aestivum* L.). Even moderate levels of HF infestation can cause severe damage (Smiley *et al.*, 2004), and up to 35% yield loss has been reported in Morocco (El Bouhssini *et al.*, 2008). Owing to the short life cycle and rapid evolution, 18 HF biotypes—A through O, GP, vH9 and vH13—have been reported based on their ability to infect genes *H3*, *H5*, *H6*, *H7H8*, *H9*, and *H13* (Formusoh *et al.*, 1996; Ratcliffe & Hatchett, 1997; Zantoko & Shukle, 1997; Ratcliffe *et al.*, 2000). It is challenging to control HF using conventional methods, such as insecticidal seed treatment and late planting (Zelarayan *et al.*, 1991; Buntin *et al.*, 1992). The rapid evolution of HF biotypes resulting in the breakdown of resistance increases the difficulty of providing durable resistance.

Li *et al.* (2013) and Tan *et al.* (2017) reviewed research on the 34 HF resistance genes, *H1* through *H34*, that have been identified to date. Only 23 have been mapped to any specific chromosome. First mapping of HF resistance gene, *H6*, was done using monosomic analysis and mapped the gene to chromosome 5A (Gallun & Patterson, 1977). Other researchers later showed that *H3, H9* and *H10* were also linked to *H6* (Carlson *et al.*, 1978; Stebbins *et al.*, 1982; Ohm *et al.*, 1995). However, further studies mapped *H3, H6, H9, H10* and 11 other genes in a cluster on chromosome 1AS, which was different from their originally mapped locations (Kong *et al.*, 2005; Liu *et al.*, 2005a; Liu *et al.*, 2005b; Kong *et al.*, 2008). HF resistance genes in clusters on the same chromosome could likely be alleles of the same gene or members of a single gene family (Liu *et al.*, 2005a; Liu *et al.*, 2005b). Of the mapped genes, 15 have been located on the A-genome (Liu *et al.*, 2005a; Liu *et al.*, 2005b), one each on 2B (Amri *et al.*, 1990) and 5B (Williams *et al.*, 2003), and six on the D-genome chromosomes 1D, 3D, 4D and 6D (reviewed in

Li *et al.* (2013)). Most of the HF resistance genes have been introgressed from *Triticum turgidum* L. and *Aegilops tauschii* L (reviewed in Li *et al.* (2013)). However, rye (*Secale cereale* L.) has also contributed two HF resistance genes to wheat (Friebe *et al.*, 1996).

In a study conducted by Cambron *et al.* (2010), only five genes—*H12, H18, H24, H25,* and *H26*—out of 21 tested, provided resistance against 20 HF populations across the southeastern USA. Such diversity and virulence in the HF populations indicates the potential for HF biotypes to rapidly shift and overcome single gene resistance (McDonald *et al.*, 2014). Even though HF is an insect pest, it is considered a plant pathogen that injects effector proteins in the host system and elicit R-gene mediated response (Stuart *et al*., 2012). Multiple studies have shown HF resistance in wheat follows a typical gene-for-gene model that exerts a strong selection pressure on the HF populations (El Bouhssini *et al.*, 2008; McDonald *et al.*, 2014). Due to such rapid evolution of HF populations, new mapped resistance genes are needed, which can be pyramided and rapidly deployed in wheat breeding programs. However, the dearth of linked molecular markers for marker assisted breeding hinders the utilization of existing genes in the crop improvement for insect resistance.

With the advancements in sequencing technologies, genotyping-by-sequencing (GBS) approaches have become a mainstay for genomics assisted breeding (Elshire *et al.*, 2011; Poland *et al.*, 2012). GBS can generate thousands of molecular markers without requiring prior SNP information. Utilizing thousands of markers not only allows high-resolution mapping, but can pinpoint and delimit introgression segments. Another innovative technology is targeted sequencing of only a portion of the genome using sequence captures (Cronn *et al.*, 2012). One such modified method is RenSeq that uses RNA bait libraries to capture only NBS-LRR

sequences (Jupe *et al.*, 2013; Steuernagel *et al.*, 2016). In some cases, targeted sequencing can help narrow down to a handful of candidate genes to facilitate cloning a gene of interest.

In this study, we genetically mapped three previously known genes: *H5, H10* and *H13* (Patterson *et al.*, 1994), of which, only *H13* has been definitively mapped on the distal end of chromosome 6DS (Gill *et al.*, 1987; Liu *et al.*, 2005c). *H5* was originally reported on chromosome 5A, and then later updated to 1A based on monosomic analysis (Gallun & Patterson, 1977; Roberts & Gallum, 1984). There is also a discrepancy in the position reported for *H10*, which was originally mapped to chromosome 5A (Stebbins *et al.*, 1982), and was later mapped to chromosome 1AS (Liu *et al.*, 2005b). In addition to these known genes, we also mapped a resistance gene introgressed from *Ae. tauschii* accession *KU2147* (El Bouhssini *et al.*, 2008). Direct crossing of winter wheat cultivar 'Overley' with *Ae. tauschii KU2147*, followed by advanced backcrossing, was used to introgressed the resistance into and elite background (Gill & Raupp, 1986).

Developing wheat cultivars with genetic resistance against HF is the best option to control this insect, economically and ecologically (El Bouhssini *et al.*, 2008). Moreover, Anderson *et al.* (2011) showed that HF resistance can be developed in wheat without a fitness cost. We hypothesized that (i) all genes in our populations were segregating as single dominant genes, (ii) introgression from *Ae. tauschii* carries NBS-LRR genes underlying HF resistance, and (iii) HF resistance gene from *Ae. tauschii* did not have any yield penalty due to linkage drag. Therefore, the main objectives of this study were to: (i) genetically map three known HF resistance genes and place them on physical positions using genotyping-by-sequencing, (ii) map a HF resistance gene transferred from *Ae. tauschii* accession *KU2147* to hard red winter wheat cultivar 'Overley' and estimate the introgression size, (iii) conduct yield tests to detect any yield

penalty associated with the *Ae. tauschii* introgression, and (iv) find candidate genes in the *KU2147* introgressed region using RenSeq. In addition, Fam1 was segregating for 'Overley' like green and *Ae. tauschii* like neon-green color. Studies in other species have reported the effect of foliage color on the insect feeding preference (Jayaraj & Uthamassamy, 1990; Cramer *et al*., 2014; Green *et al*., 2015). Therefore, we also investigated if HF resistance was affected by the plant color under field conditions.

Materials and methods

$F_{2:3}$ lines development

Erin, Joy and Molly have been released as HF resistant germplasm isolines to the cultivar 'Newton' with resistance genes *H5*, *H10* and *H13*, respectively (Patterson *et al.*, 1994). Using these resistant isolines, we developed HF mapping populations by crossing them to the recurrent parent 'Newton'. Populations hereafter referred to as H5-EN, H10-JN and H13-*MN*, respectively. This was followed by two generations of selfing to get $F_{2:3}$ lines. In addition, Molly was also crossed with 'Overley' to develop H13-MO population. Seeds for the germplasm isolines were received from USDA-ARS, Manhattan, KS, USA. Pedigrees of these populations are listed in Supplementary Table D.1.

$BC_3F_{2:3}$ lines development

One $BC_3F_{2:3}$ mapping population was developed by direct crossing hexaploid hard red winter wheat cultivar 'Overley' with diploid *Ae. tauschii* accession *KU2147* following the approach of Gill and Raupp (1987). This was followed by three backcrosses and two rounds of selfing to develop $BC_3F_{2:3}$ lines. While screening BC_1 generation for HF resistance, two resistant plants were identified and propagated separately to $BC_3F_{2:3}$ to develop two families, hereafter

noted as Fam1 and Fam2. Pedigree information for this population is listed in Supplementary Table D.1.

Phenotypic screening

Phenotypic screening for all populations was conducted in a similar fashion (Liu *et al.*, 2005a; Garcés-Carrera *et al.*, 2014). Lines were grown in 4-inch deep plastic trays in the USDA-ARS greenhouse at Kansas State University, Manhattan, KS, USA. Greenhouse temperature was maintained at 20˚C for the whole screening season. Trays were filled with 3-inches of soil and divided into 24 furrows accommodating up to 20 equally spaced experimental lines and four checks—Karl 92 (susceptible), Carol (*H3*; moderately resistant), Caldwell (*H6*; resistant) and Molly (*H13*; resistant). Fourteen trays were laid out in a set in an augmented design where each tray consisted of same four checks. Twenty seeds were sown for each line and checks. Immediately after sowing, the seed was covered with a very thin layer of mixture of 3:1 soil and MetroMix, and irrigated till saturation.

At seven to eight days after sowing, at the one to two leaf stage, seedlings were infested with HF Great Plains (GP) biotype from stock maintained at USDA-ARS facility in Manhattan, KS. Trays were covered with white cheesecloth nets to keep high humidity and restrict the escape of the adult flies. Periodic egg count per leaf was conducted, and flies were removed once an average of 10-15 eggs per leaf were observed.

Phenotypic scoring was performed 15-17 days after infestation. Lines were scored as homogeneously resistant (phenotype R; genotype RR) lines, homogeneously susceptible (S; rr) lines, or heterogeneous (H; Rr) lines, if they had 100% resistant, 100% susceptible, and mixture of resistant and susceptible plants, respectively. Resistant lines were scored again 3-4 days after initial scoring for confirmation. Not all lines germinated completely, therefore the lines with less

than 16 plants were not scored to avoid scoring a heterogeneous line as R. Lines with only a single susceptible plant, out of 20 plants, were also discarded because of possible seed mixture.

Fam1 was also segregating for plant color with wheat-like green and *Ae. tauschii*-like neon-green color, and therefore was subjected to visual color phenotyping. Color phenotyping was done at adult plant stage in the field. Lines were planted in head rows with 20-25 plants each. Color data for RR, Rr and rr lines were collected separately, and to test the association, Chi-square test was run at alpha level of 0.05.

Field testing in yield plots

To test if there was any yield penalty associated with the gene from *KU2147*, 55 homozygous resistant and 55 homozygous susceptible $BC_3F_{2:3}$ lines were planted in 6-row yield plots. Lines were planted in augmented design in 10 blocks with three checks—'Everest', 'SY Flint', and 'Overley'—in each block. Sixty grams of seed was planted for each plot. Data were recorded for plant height after plants fully matured, grain weight, moisture content at harvesting, test weight, and relative days to heading from January 1, 2017. Grain weights were adjusted for the plots with less than six full rows. Welch two sample t-test was performed for each trait between resistant and susceptible lines to detect any significant differences.

Tissue collection and genotyping-by-sequencing

Tissue from six to eight plants for each of the homozygous R and S lines was collected in 96-well plates for DNA sequencing. Tissue was lyophilized for 48 hours before grinding. DNA was extracted using Qiagen BioSprint 96 DNA Plant Kit (QIAGEN, Hilden, Germany), and quantified with Quant-iT™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA) and normalized. Genotyping-by-sequencing (GBS) libraries were prepared according to Poland *et al.* (2012) using 96-plexing with one blank control (95-plex). Libraries were

sequenced on Illumina HiSeq2000 (Illumina, San Diego, CA, USA). GBS was not performed for Fam2.

R-gene enrichment sequencing (RenSeq)

Additionally, Fam1 and Fam2 were subjected to RenSeq for targeted sequencing of NBS-LRR genes using RNA bait libraries (Jupe *et al.*, 2013; Steuernagel *et al.*, 2016). Tissue from homozygous R and S lines for both families were collected individually. DNA was extracted using Qiagen BioSprint 96 DNA Plant Kit (QIAGEN, Hilden, Germany), and quantified with Quant-iT™ PicoGreen® dsDNA Assay Kit (ThermoFisher Scientific, Waltham, MA, USA) and normalized. For each family, equimolar DNA from all the resistant lines was bulked into an R bulk, and susceptible lines into an S bulk. A total volume of 100µl with DNA concentration of 40ng/µl was used for library preparation. DNA quality was checked using 1% high resolution agarose gel electrophoresis. DNA was enriched for R-gene using small RNA baits that were complementary to the conserved NBS-LRR domains of R-genes. Libraries for sequencing were prepared with an insert size of 800bp and sequenced with 150bp paired-end sequencing on Illumina HiSeq2500.

SNP discovery, genotyping and SNP filtering

Raw GBS reads for H5-EN, H10-JN, H13-MN, and H13-MO collectively, were aligned to International Wheat Genome Sequencing Consortium Reference Sequence v1.0 (RefSeq v1.0) assembly (http://www.wheatgenome.org/) for SNP discovery using TASSEL5 GBSv2 pipeline (Glaubitz *et al.*, 2014) with default settings. Output files containing SNP information in hapmap format were processed using R statistical language (R Core Team, 2015). For each population, individuals were parsed out in separate datasets using population identifiers. For each population, SNPs with more than two alleles were discarded. SNPs were filtered for less than 20% missing

data, and minor allele frequency greater than 0.3. Similar procedure was followed for Fam1 GBS data, except that missing data threshold was adjusted to 50% to retain more SNPs. Further SNP filtering was performed by removing the SNPs that were monomorphic between parents, or heterozygous or missing in either of the two parents.

RenSeq reads for Fam1 and Fam2 were aligned with RefSeq v1.0 using Burrows-Wheeler Aligner with 'bwa mem' algorithm (Li & Durbin, 2009). Generated SAM files were sorted and processed with 'samtools' to remove duplicated tags. Finally, 'samtools mpileup' was used to call SNPs. Monomorphic SNPs that were present only due to the reference assembly were removed. Using a customized PERL script (https://github.com/liu3zhenlab/scripts/blob/master/VCF/vcf2AC.pl), counts for reference and alternate alleles were computed for each SNP site for each bulk. SNPs with any reference or alternate allele count less than five, or the total depth per bulk less than 50, were removed. From the remaining, only those SNPs, where R bulk had higher number of alternate allele and S bulk had higher number of reference allele, were used for further analyses.

Statistical analysis and mapping

For phenotypic distributions, Chi-square for goodness of fit tests were performed to determine the genetic segregation of the resistance genes. Tests were performed for each population separately to see if the samples fit into 1:2:1 segregation ratio for R:H:S lines with an alpha level of 0.05. For GBS SNP markers, single marker association tests were performed in R statistical language (R Core Team, 2015). *P*-values were computed for each marker with phenotype score as a response variable using analysis of variance (ANOVA). Contrasting to GBS SNP markers, *p*-values for RenSeq allele counts were computed using Fisher's exact test for allele counts. *P*-values were then plotted using 'qqman' package in R (Turner, 2014). SNPs

above the population specific Bonferroni's threshold, at experimental-wise alpha of 0.001, were considered significant and used to infer putative location of the resistance gene, and the introgression size of the segment from *Ae. tauschii*.

Identifying candidate genes

Significant SNPs from RenSeq data were used to demarcate the gene rich region, carrying NBS-LRR genes, within the introgression. Additional 50Kb flanking sequences on the both sides of this region were surveyed on the annotated RefSeq v1.0 assembly, and visualized using JBrowse (Skinner *et al.*, 2009) set up on IWGSC website. Gene names and relevant information was retrieved from the annotated assembly.

Allele matching

To confirm if the source germplasm lines from Patterson *et al.* (1994) were in fact isolines with 'Newton' or not, allele matching was performed as defined in Chapter 2. Complete SNP set was used for this comparison but for each germplasm line, chromosome with the gene introgression was removed from the comparison, and then compared with 'Newton'. A threshold of 95% allele matching was applied to declare if the two lines were isolines or not.

Results

Phenotypic screening for Hessian fly resistance

For $F_{2:3}$ populations, an equal number of 150 lines were planted for each of the populations. After discarding the lines based on poor germination and single susceptible plants the populations H5-EN, H10-JN, H13-MN and H13-MO had 146, 150, 113, and 87 lines, respectively, which were screened for HF resistance. For $BC_3F_{2:3}$ Fam1 and Fam2 from *KU2147*, a total number of 316 and 273 lines were screened, respectively.

Checks in all the trays and sets carried the expected phenotypic reactions to HF biotype GP. Molly (*H13*) was completely resistant, Karl92 was completely susceptible, and Carol (*H3*) showed heterogeneous phenotypes with mostly susceptible plants. Caldwell (*H6*) was homogeneously resistant in most of the trays, but showed a small number of susceptible plants in few trays.

With the exception of H5-EN, phenotypic distribution of the $F_{2:3}$ lines in each population indicated that the resistance genes are segregating as single dominant genes (Fig. 4.1). This was tested using Chi-square goodness-of-fit test at alpha value 0.05. H5-EN failed the Chi-square test with p-value < 0.0001 because of excessive number of homozygous susceptible lines (Fig. 4.1). However, H10-JN, H13-MN, and H13-MO failed to reject the null hypothesis that the resistance is conferred by single dominant gene (p-value > 0.05). Similarly, phenotypic distribution for both Fam1 and Fam2 fits 1:2:1 ratio (Fig. 4.1), which indicates that the HF resistance gene from *KU2147* is also a single dominant gene providing immunity against HF biotype GP.

Fam1 was also phenotyped for foliage color, under field conditions, as either homogeneously wheat green, homogeneously *Ae. tauschii* neon-green, or segregating for both colors. Phenotypic distribution for the plant color also fits 1:2:1 ratio, indicating that the plant color is also possible controlled by single dominant gene (left panel, Fig. D.1). Chi-square test for the independence of plant color and HF resistance indicated that plant color and HF resistance are under independent genetic control (right panel, Fig. D.1; p-value = 0.5276).

Field trial to test fitness cost of gene from *KU2147*

Yield tests were conducted to investigate if the introgressed genetic resistance from *Ae. tauschii* had any associated agronomic penalty. Paired t-test was conducted and revealed no significant differences between homozygous resistant and homozygous susceptible lines

performance for any of the recorded traits. Adjusted grain weight showed non-significant fluctuation between both groups (Fig. 4.2). Similar pattern was observed for all other traits (Fig. D.2). Adjusted grain weight, grain moisture and plant height showed near normal phenotypic distribution, whereas test weight showed left skewed distribution and days to heading showed categorical distribution.

SNP discovery and genotyping data

Raw GBS reads from all populations were aligned against RefSeq v1.0, and putative SNPs were discovered. Initially, 39,671 putative SNPs were identified for all the $F_{2:3}$ populations collectively. Per SNP missing data ranged from 0-98% across all the populations. To retain high-quality SNPs, each population was filtered to remove the SNPs with more than 20% missing data and MAF less than 0.3, which drastically reduced the final number of SNPs across the populations (Table 1). For Fam1, initially 17,189 SNPs were discovered, which were filtered to retain only SNPs with greater than 0.3 MAF and less than 50% missing data. Finally, all populations had comparable number of SNPs.

Statistical analysis and gene mapping with GBS data

All the populations with the known genes showed one clear single peak with multiple significant SNPs at the same region (Fig. 4.3). However, in H5-EN, there were two significant SNPs, one each on chromosomes 1B and 2B, above the Bonferroni threshold. Gene *H5* was mapped on the tip of the short arm of chromosome 6B with an estimated introgression size of 5.96 Mb. Gene *H10* was mapped on the distal part of the short arm of chromosome 6D with an estimated introgression of size 5.01 Mb.

From the known genes in this study, only *H13* has been definitively mapped before. Therefore, *H13* was used as a control for the GBS approach in this study. *H13* segregating

populations were developed in two different backgrounds, 'Newton' and 'Overley'. Using both populations, *H13* was unanimously mapped at the distal end of short arm of chromosome 6D, in both backgrounds, with an average estimated introgression of 6.97 Mb (Fig. 4.3). When plotted together, we found that genes *H10* from H10-JN and *H13* from H13-MN and H13-MO actually mapped at the exact same location on the chromosome 6D with overlapping SNPs (Fig. 4.4).

Apart from the known genes in isoline populations, we also mapped an introgressed gene, *H2147*, in $BC_3F_{2:3}$ Fam1 from *Ae. tauschii* accession *KU2147*. With only one significant SNP mapping on a different chromosome 3B, all significant SNPs mapped on 3D (panel A, Fig. 4.5). Filtering only the significant SNPs, the introgressions size was estimated to be 6.35 Mb and located towards the distal end of long arm of chromosome 3D.

R-gene enrichment sequencing and identification of candidate genes

A targeted exome capture technique, RenSeq, was used to facilitate the discovery of candidate HF resistance genes. RenSeq involves capturing only NBS-LRR genes using RNA bait libraries, followed by high-throughput-sequencing. RenSeq generated more than 4.6 million SNPs between resistant and susceptible bulks. With the filtering criteria described in the methods, final high-quality SNP counts for Fam1 and Fam2 were 4,113 and 4,398, respectively. For both families, almost all the significant SNPs mapped on chromosome 3D (Fig. 4.5). Only a few SNPs mapped to other homoeologous 'group 3' chromosomes. Fam2 mapping results confirmed the results of Fam1. Using only significant SNPs, the gene was putatively mapped on the distal end of chromosome 3DL, possibly between or around 575.48-575.54 Mb.

Annotated assembly of RefSeq v1.0 was surveyed to find candidate genes in the above-demarcated region and flanking 50Kb sequences on both sides. Seven NBS-LRR genes were identified in this region: TraesCS3D01G473700, TraesCS3D01G473800,

TraesCS3D01G473900, TraesCS3D01G474000, TraesCS3D01G474100, TraesCS3D01G474200, and TraesCS3D01G474300. All being NBS-LRR genes, these seven genes represent good candidates for the underlying resistance gene on chromosome 3D.

Germplasm isolines are not isolines

Due to the unexpected high number of SNPs observed in all isoline populations, we tested if the germplasm isolines were in fact isolines with 'Newton'. Allele matching was performed and pairwise identity coefficient was computed after removing the chromosome carrying the gene of interest. Identity coefficients ranged from 0.8566 (85.66%) for 'Newton'-Joy to 0.8834 (88.34%) for 'Newton'-Erin. No isolines reached our 0.95 (95%) threshold, indicating that the germplasm lines were not isolines.

Discussion

Hessian fly is a devastating pest of wheat worldwide that infects plants as a pathogen by injecting effector proteins and eliciting NBS-LRR genes mediated defense response from the host (Stuart *et al*., 2012). This prior knowledge guided our decisions to design the experiment to map genes and find candidate genes for the newly introgressed gene. Multiple mapping populations were developed to map one new and three known genes providing resistance against HF biotype GP. New genetic resistance was derived from *Ae. tauschii*, a wild relative of wheat, and known resistance genes were mapped from the germplasm lines released by Patterson *et al.* (1994). Four $F_{2:3}$ and one $BC_3F_{2:3}$ (two families) populations were developed. Our success in mapping all the genes at a single chromosomal location provides evidence of sufficient population sizes for mapping single dominant genes.

Controls used in the study behaved as expected with Molly being homogeneously resistant, Karl 92 homogeneously susceptible, and Carol the mixture of both R and S. However,

Caldwell showed a few instances of discrepancy where in few trays we found susceptible plants, which could possibly be due to either the seed mixture during handling or planting, or due to HF biotype shift within biotype GP. Presence of virulent flies at low frequency could also explain the susceptibility of few Caldwell plants. Overall, expected phenotypic distribution of controls provides an evidence that the experimental conditions were conducive for this mapping study.

All mapping populations were segregating for a single HF resistance gene, and segregated as expected to fit Mendelian 1:2:1 phenotypic ratio (Fig. 4.1). The only exception was H5-EN population that did not follow a typical F_2 segregation ratio (Fig. 4.1). This can possibly be attributed to a small population size (n=146). However, as evident in the other populations with similar population sizes, this is unlikely, therefore some other factor(s) might have caused this inconsistency. When developed and released as a HF resistant germplasm, Erin was not tested against GP biotype, but only against biotypes B, C, D and L (Patterson *et al.*, 1994). In that study, Erin was resistant to all tested biotypes except biotype L. Therefore, it is possible that *H5* gene in H5-EN does not provide complete resistance against biotype GP. Another gene, *H34*, is also reported to have partial resistance to HF biotype GP (Li *et al.*, 2013). However, *H34* was reported as a major effect QTL rather than a single gene. Another reason for the partial resistance by *H5* could be the temperature fluctuations in the greenhouse. Although the greenhouse was maintained at 20˚C, there is a possibility that the temperature fluctuated and caused the *H5* gene to fail. Tyler and Hatchett (1983) had also reported the loss of resistance from *Ae. tauschii* at higher temperatures.

Other plant phenotypes, such as plant color, volatile compounds, or lipid content, usually affect insect feeding behavior in plants (Eigenbrode & Espelie, 1995; Bruce & Pickett, 2011). Since Fam1 was segregating for wheat like green color and *Ae. tauschii* like neon green color,

we phenotyped this family for plant color, and attempted to see if there was any correlation between plant color and HF resistance. Chi-square test for the independence of two factors failed to reject the null that two factors, plant color and HF resistance, were independent of each other. Therefore, we concluded that in Fam1, plant color did not affect HF resistance. However, there might be other factors, such as *Ae. tauschii* like pubescence, that can affect resistance.

IWGSC RefSeq v1.0 was used as a reference to align the GBS reads and find physical position of SNPs. SNPs with more than two alleles were discarded because multi-allelic SNPs are not expected in a bi-parental population. The alleles in a bi-parental population should segregate at intermediate frequencies, therefore, only SNPs with MAF greater than 0.3 were used in the analyses. However, after filtering on these criteria, higher than expected number of SNPs were discovered for isoline populations, whereas, lower number of SNPs were expected based on the population design. Germplasm lines Erin, Joy and Molly were released as isolines to a popular wheat cultivar 'Newton'. However, allele matching analysis revealed that none of these germplasm lines matched with 'Newton' with 95% or greater identity, which could also explain why we observed higher than expected number of SNPs.

Single marker association analysis for each population revealed that each gene was mapped to a single chromosomal location. Single marker analysis over interval mapping is particularly appropriate because of its computational efficiency and the availability of genome-wide dense markers. In 'Newton' background, *H5* was mapped on chromosome 6B, and *H10* and *H13* were mapped on 6D. As an independent validation, *H13* was mapped to the same location in 'Overley' background, which confirms that this method is robust for future mapping of other genes. A few markers were mapped on other chromosomes at random positions across the

genome, which is possibly due to the mapping of short GBS reads at wrong genomic positions due to sequence similarity.

Initially, *H5* was mapped on chromosome 5A, and then later updated to 1A based on monosomic analysis (Gallun & Patterson, 1977; Roberts & Gallum, 1984). In wheat, monosomic analysis is performed by crossing a line carrying normal 21 pairs of chromosomes (euploid) with 21 lines lacking one particular chromosome (aneuploids) (Sears, 1953; Singh, 1967). Resulting F1 hybrids are hemizygous for the monosomic chromosome, and are selfed to produce F2 progeny. F2 progeny for each cross should result in typical 3:1 phenotypic ratio, except for the chromosome carrying the gene of interest. Roberts & Gallum (1984) found that the crosses with monosomic lines of chromosome 1A, 2A and 3D were violating 3:1 phenotypic ratio, and they concluded that the gene *H5* was present on 1A based on 97:3 ratio for this cross. However, their population sizes were not adequate to have reached that conclusion definitively. Therefore, threre is still a discrepancy regarding the position of *H5*, which we plan to investigate further.

There was another discrepancy regarding the positon of gene *H10* that was mapped on chromosome 6DS, however, it was previously reported on chromosome 5A, but was later updated to 1AS (Liu *et al.*, 2005b). In our study, when plotted together, genes *H10* and *H13* genes mapped on the exact same genomic location of chromosome 6DS (Fig. 4.4). This points to the possibility of pollen contamination from *H13* resistant source parent "Molly" during population development, and we could be mapping *H13* gene instead of *H10*. Another possibility is that the region associated with *H10* on 1AS was missing in the reference genome we used but had the sequence similarity to 6DS. This could also result in *H10* mapping to chromosome 6DS. Since typical HF resistance genes are NBS-LRR type genes, they generally tend to occur in clusters in wheat (Liu *et al.*, 2005a; Liu *et al.*, 2005b), and other plants species, such as potato

(Park *et al.*, 2005), sunflower (Qi *et al.*, 2011), and grapevine (Coleman *et al.*, 2009). If the *H10* mapping in our study is accurate, it is probable that *H10* and *H13* are allelic or in a gene family. However, at this point no allelism tests can distinguish them and we do not have sufficient population sizes to fine map the genes to the resolution needed to reach any conclusion. Further investigation is required to accurately conclude the *H10* position.

The new gene, *H2147,* from *Ae. tauschii* was mapped on the long arm of chromosome 3D in Fam1. The majority of the SNPs were mapped on chromosome 3D but a few mapped on 3A and 3B. Wheat is an allohexaploid species with two homoeologous chromosomes from the A and B genomes for each D genome chromosome. One single marker showing up on 3A is likely due to misalignment of GBS reads due to sequence similarity between the three genomes. Using only significant markers this gene was mapped in a 6.35 Mb introgression from *Ae. tauschii*.

As reported by Anderson *et al.* (2011), we also hypothesized that HF resistance can be developed in wheat without a yield penalty. To test this hypothesis for *H2147*, we planted 55 lines from both the homozygous resistant and susceptible classes in yield plots. Comparison of the data for five different traits confirmed that there was no significant difference among the resistant lines and susceptible lines. Moreover, both resistant and susceptible lines did not differ from the recurrent parent 'Overley'. Two unrelated, but similar, control cultivars, 'Everest' and 'SY Flint', were also phenotyped as expected and did not differ from the tested lines. SY Flint carries a Hessian fly resistance gene transferred from *Triticum turgidum* ssp. *dicoccum* via KS99WGRC42 (Brown-Guedira *et al.*, 1999). These findings present a strong evidence that the gene *H2147* has no yield penalty or detrimental effect on any other measured traits in the lines carrying it.

After mapping a gene or a QTL, generally the next step is to delineate the genomic region to a smaller level, and identify candidate genes. This usually involves fine mapping the gene by developing a large NIL or F_2 population to reduce the size of introgression using natural recombination. As a first step, we applied a modified exome capture method, known as RenSeq, which involved capturing and sequencing only the NBS-LRR genes to identify candidate genes. There is good evidence that the HF resistance genes belong to a typical R-gene family with NBS-LRR domains based on the ability of HF to elicit R-genes mediated host response by injecting effector proteins (Stuart *et al.*, 2012). With this hypothesis, we used the RNA bait libraries to capture only NBS-LRR genes from the genomic DNA of resistant and susceptible bulks from Fam1 and Fam2 (Steuernagel *et al.*, 2016). Sequencing at higher depth and analyzing the RenSeq data, we were able to map this gene on the distal end of chromosome 3D long arm, consistent with the GBS results of Fam1. Since RenSeq generates high density of markers in genic regions, we were able to identify significant SNPs and could identify putative candidate genes for *H2147* in the introgression.

Surveying the annotated RefSeq v1.0 assembly for the candidate genes in the introgression, we were able to identify seven candidate genes in this region. All seven genes were disease resistance genes with conserved NBS-LRR domains. All of these genes, except one, were predicted in wheat based on the gene models. However, TraesCS3D01G474100 is the smallest predicted gene (513bp) in this region based on its sequence similarity to *Arabidopsis thaliana* NBS-LRR gene. Next set of NBS-LRR type genes were located 293Kb upstream of TraesCS3D01G473700 and 1.7Mb downstream of TraesCS3D01G474300, which suggests that these seven genes could be good candidates for future cloning efforts.

In conclusion, we mapped three known genes and one gene introgressed from *Ae. tauschii* providing resistance against biotype GP. HF damage is a serious threat to wheat growing regions not only in the US but worldwide. Developing wheat cultivars with multiple genes stacked in a single wheat cultivar would be a better option to control broad range of HF biotypes, and we anticipate that this study would be very beneficial to facilitate this breeding goal for strengthening the future food security.

Acknowledgements

I would like to thank Dr. Ming-Shun Chen for providing greenhouse and Hessian fly for conducting phenotypic screening, and Dr. Sanzhen Liu for the discussion and guidance on doing RenSeq data analysis, and providing me with the script to compute reference and alternate alleles. I would also like to thank Dr. Brande Wulff (John Innes Center, Norwich, UK) for providing us with the RNA bait libraries for RenSeq enrichment. Also, we would like to acknowledge our funding partner Kansas Wheat Commission for funding this project. This study was conducted under the auspices of the Wheat Genetics Resource Center (WGRC) Industry/University Collaborative Research Center (I/UCRC) through support of industry partners and partially funded by NSF grant contract (IIP-1338897).

References

Amri A, Cox T, Gill B, Hatchett J. 1990. Chromosomal location of the Hessian fly resistance gene H20 in 'Jori'durum wheat. *Journal of Heredity* **81**(1): 71-72.

Anderson KM, Kang Q, Reber J, Harris MO. 2011. No fitness cost for wheat's H gene-mediated resistance to Hessian fly (Diptera: Cecidomyiidae). *Journal of Economic Entomology* **104**(4): 1393-1405.

Brown-Guedira GL, Hatchett JH, Cox TS, Sears RG, Owuoche JO, Gill BS. 1999. Notice of Release. The Kansas Agricultural Experiment Station, Manhattan, Kansas.

Bruce TJ, Pickett JA. 2011. Perception of plant volatile blends by herbivorous insects–finding the right mix. *Phytochemistry* **72**(13): 1605-1611.

Buntin G, Ott S, Johnson J. 1992. Integration of plant resistance, insecticides, and planting date for management of the Hessian fly (Diptera: Cecidomyiidae) in winter wheat. *Journal of Economic Entomology* **85**(2): 530-538.

Cambron SE, Buntin GD, Weisz R, Holland JD, Flanders KL, Schemerhorn BJ, Shukle RH. 2010. Virulence in Hessian fly (Diptera: Cecidomyiidae) field collections from the southeastern United States to 21 resistance genes in wheat. *Journal of Economic Entomology* **103**(6): 2229-2235.

Carlson S, Patterson F, Gallun RL. 1978. Inheritance of Resistance to Hessian Fly Derived From *Triticum turgidum* L. *Crop Science* **18**(6): 1011-1014.

Coleman C, Copetti D, Cipriani G, Hoffmann S, Kozma P, Kovács L, Morgante M, Testolin R, Di Gaspero G. 2009. The powdery mildew resistance gene REN1 co-segregates with an NBS-LRR gene cluster in two Central Asian grapevines. *BMC genetics* **10**(1): 89.

Cramer CS, Singh N, Kamal N, Pappu HR. 2014. Screening Onion Plant Introduction Accessions for Tolerance to Onion Thrips and Iris Yellow Spot. *HortScience* **49**(10): 1253-1261.

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* **99**(2): 291-311.

Eigenbrode SD, Espelie KE. 1995. Effects of plant epicuticular lipids on insect herbivores. *Annual review of entomology* **40**(1): 171-194.

El Bouhssini M, Nachit MM, Valkoun J, Abdalla O, Rihawi F. 2008. Sources of resistance to Hessian fly (Diptera: Cecidomyiidae) in Syria identified among *Aegilops* species and synthetic derived bread wheat lines. *Genetic Resources and Crop Evolution* **55**(8): 1215-1219.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**(5): e19379.

Formusoh E, Hatchett J, Black W, Stuart J. 1996. Sex-linked inheritance of virulence against wheat resistance gene H 9 in the Hessian fly (Diptera: Cecidomyiidae). *Annals of the Entomological Society of America* **89**(3): 428-434.

Friebe B, Jiang J, Raupp WJ, McIntosh RA, Gill BS. 1996. Characterization of wheat-alien translocations conferring resistance to diseases and pests: Current status. *Euphytica* **91**(1): 59-87.

Gallun R, Patterson F. 1977. Monosomic analysis of wheat for resistance to Hessian fly. *Journal of Heredity* **68**(4): 223-226.

Garcés-Carrera S, Knutson A, Wang H, Giles KL, Huang F, Whitworth RJ, Smith CM, Chen M-S. 2014. Virulence and biotype analyses of Hessian fly (Diptera: Cecidomyiidae) populations from Texas, Louisiana, and Oklahoma. *Journal of Economic Entomology* **107**(1): 417-423.

Gill B, Hatchett J, Raupp W. 1987. Chromosomal mapping of Hessian fly-resistance gene *H13* in the D genome of wheat. *Journal of Heredity* **78**(2): 97-100.

Gill BS. 1986. Resistance in *Aegilops* squarrosa to Wheat Leaf Rust, Wheat Powdery Mildew, Greenbug, and Hessian Fly. *Plant Disease* **70**(6): 553-553.

Gill BS, Raupp WJ. 1987. Direct Genetic Transfers from *Aegilops*-Squarrosa L to Hexaploid Wheat. *Crop Science* **27**(3): 445-450.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**(2): e90346.

Green JP, Foster R, Wilkins L, Osorio D, Hartley SE. 2015. Leaf colour as a signal of chemical defence to insect herbivores in wild cabbage (Brassica oleracea). *PLoS One* **10**(9): e0136884.

Jayaraj S, Uthamassamy S. 1990. Aspects of insect resistance in crop plants. *Proceedings: Animal Sciences* **99**(3): 211-224.

Jupe F, Witek K, Verweij W, Śliwka J, Pritchard L, Etherington GJ, Maclean D, Cock PJ, Leggett RM, Bryan GJ, *et al.* 2013. Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *The Plant Journal* **76**(3): 530-544.

Kong L, Cambron SE, Ohm HW. 2008. Hessian fly resistance genes H16 and H17 are mapped to a resistance gene cluster in the distal region of chromosome 1AS in wheat. *Molecular Breeding* **21**(2): 183-194.

Kong L, Ohm H, Cambron S, Williams C. 2005. Molecular mapping determines that Hessian fly resistance gene H9 is located on chromosome 1A of wheat. *Plant Breeding* **124**(6): 525-531.

Li C, Chen M, Chao S, Yu J, Bai G. 2013. Identification of a novel gene, H34, in wheat using recombinant inbred lines and single nucleotide polymorphism markers. *Theoretical and Applied Genetics* **126**(8): 2065-2071.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Liu X, Brown-Guedira G, Hatchett J, Owuoche J, Chen M-S. 2005a. Genetic characterization and molecular mapping of a Hessian fly-resistance gene transferred from T. *turgidum* ssp. dicoccum to common wheat. *Theoretical and Applied Genetics* **111**(7): 1308-1315.

Liu X, Fritz A, Reese J, Wilde G, Gill B, Chen M-S. 2005b. H9, *H10*, and H11 compose a cluster of Hessian fly-resistance genes in the distal gene-rich region of wheat chromosome 1AS. *Theoretical and Applied Genetics* **110**(8): 1473-1480.

Liu X, Gill B, Chen M-S. 2005c. Hessian fly resistance gene *H13* is mapped to a distal cluster of resistance genes in chromosome 6DS of wheat. *Theoretical and Applied Genetics* **111**(2): 243-249.

McDonald MJ, Ohm HW, Rinehart KD, Giovanini MP, Williams CE. 2014. H33: A Wheat Gene Providing Hessian Fly Resistance for the Southeastern United States. *Crop Science* **54**(5): 2045-2045.

Ohm H, Sharma H, Patterson F, Ratcliffe R, Obanni M. 1995. Linkage relationships among genes on wheat chromosome 5A that condition resistance to Hessian fly. *Crop Science* **35**(6): 1603-1607.

Park T-H, Gros J, Sikkema A, Vleeshouwers VG, Muskens M, Allefs S, Jacobsen E, Visser RG, van der Vossen EA. 2005. The late blight resistance locus Rpi-blb3 from Solanum bulbocastanum belongs to a major late blight R gene cluster on chromosome 4 of potato. *Molecular Plant-Microbe Interactions* **18**(7): 722-729.

Patterson FL, Maas FB, Foster JE, Ratcliffe RH, Cambron S, Safranski G, Taylor PL, Ohm HW. 1994. Registration of Eight Hessian Fly Resistant Common Winter Wheat Germplasm Lines (Carol, Erin, Flynn, Iris, Joy, Karen, Lola, and Molly). *Crop Science* **34**(1): 315-315.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* **7**(2): e32253.

Qi L, Hulke B, Vick B, Gulya T. 2011. Molecular mapping of the rust resistance gene R4 to a large NBS-LRR cluster on linkage group 13 of sunflower. *Theoretical and Applied Genetics* **123**(2): 351-358.

R Core Team. 2015. R: A language and environment for statistical computing: R Foundation for Statistical Computing.

Ratcliffe RH, Cambron SE, Flanders KL, Bosque-Perez NA, Clement SL, Ohm HW. 2000. Biotype composition of Hessian fly (Diptera: Cecidomyiidae) populations from the southeastern, midwestern, and northwestern United States and virulence to resistance genes in wheat. *Journal of Economic Entomology* **93**(4): 1319-1328.

Ratcliffe RH, Hatchett JH. 1997. Biology and genetics of the Hessian fly and resistance in wheat. *New Developments in Entomology. Research Signpost, Scientific Information Guild, Trivandrum, India*.

Roberts JJ, Gallum RL. 1984. Chromosome location of the *H5* gene for resistance to the Hessian fly in wheat. *Journal of Heredity* **75**(2): 147-148.

Sears ER. 1953. Nullisomic analysis in common wheat. *The American Naturalist* **87**: 245-252.

Singh MP. 1967. Monosomic analysis in wheat. *Heredity* **22**, 591–596.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. *Genome Res* **19**(9): 1630-1638.

Smiley RW, Gourlie JA, Whittaker RG, Easley SA, Kidwell KK. 2004. Economic impact of Hessian fly (Diptera: Cecidomyiidae) on spring wheat in Oregon and additive yield losses with Fusarium crown rot and lesion nematode. *J Economic Entomology* **97**(2): 397-408.

Stebbins NB, Patterson F, Gallun R. 1982. Interrelationships among wheat genes H3, H6, H9, and *H10* for Hessian fly resistance. *Crop Science* **22**(5): 1029-1032.

Steuernagel B, Periyannan SK, Hernández-Pinzón I, Witek K, Rouse MN, Yu G, Hatta A, Ayliffe M, Bariana H, Jones JD. 2016. Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nature Biotechnology* **34**(6): 652-655.

Stuart JJ, Chen M-S, Shukle R, Harris MO. 2012. Gall Midges (Hessian Flies) as Plant Pathogens. Annu Rev Phytopathol **50**:339–357.

Tan CT, Yu H, Yang Y, Xu X, Chen M, Rudd JC, Xue Q, Ibrahim AMH, Garza L, Wang S, *et al.* 2017. Development and validation of KASP markers for the greenbug resistance gene Gb7 and the Hessian fly resistance gene H32 in wheat. *Theor Appl Genet*: 1-18.

Turner SD. 2014. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *BioRxiv*: 005165.

Tyler J, Hatchett J. 1983. Temperature Influence on Expression of Resistance to Hessian Fly (Diptera: Cecidomyiidae) in Wheat Derived from *Triticum* tauschii 1. *Journal of Economic Entomology* **76**(2): 323-326.

Williams C, Collier C, Sardesai N, Ohm H, Cambron S. 2003. Phenotypic assessment and mapped markers for H31, a new wheat gene conferring resistance to Hessian fly (Diptera: Cecidomyiidae). *Theoretical and Applied Genetics* **107**(8): 1516-1523.

Zantoko L, Shukle R. 1997. Genetics of virulence in the Hessian fly to resistance gene *H13* in wheat. *Journal of Heredity* **88**(2): 120-123.

Zelarayan E, Buntin G, Johnson J, Bruckner P, Raymer P. 1991. Integrated management for Hessian fly in Triticale. *Journal of production agriculture* **4**(4): 629-633.

Table 4.1. SNP summary for HF mapping populations.

Population code	***Total SNPs***	***Bi-allelic SNPs***	***SNPs after filtering***
H5-*EN*	39,671	29,261	5,629
H10-*JN*		32,246	5,491
H13-*MN*		29,210	5,518
H13-*MO*		34,644	4,660
Fam1	17,189	16,847	4,741

Table 4.2. Percent identity coefficients of germplasm isoline donors with recurrent parent 'Newton'.

SNPs from all the chromosomes were compared except the chromosomes with mapped gene for each donor.

Germplasm isoline	***Percent identity with 'Newton'***	***Number of SNP comparisons***
Erin	0.8834	13,405
Joy	0.8566	14,097
Molly	0.8782	14,437

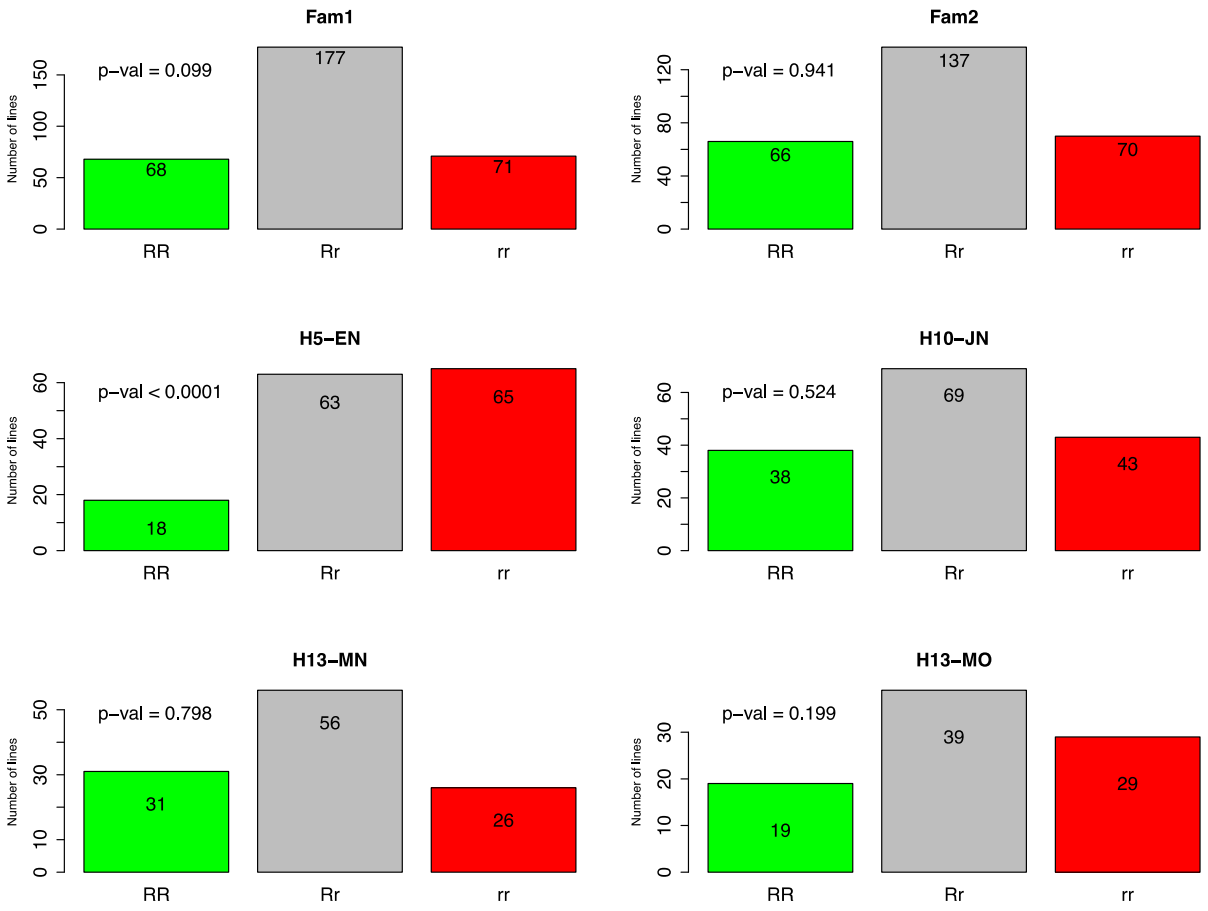

Figure 4.1. Phenotypic distribution of homogeneous resistant (phenotype R; genotype RR), heterogeneous (H; Rr) and homogeneous susceptible (S; rr) lines for all the populations.

***P*-values from Chi-square test are shown on the top left of each panel. Values inside the bars represent the number of lines observed for each genotypic group. *P*-values greater than 0.05 failed to reject the null hypothesis that the HF resistance is controlled by a single dominant gene.**

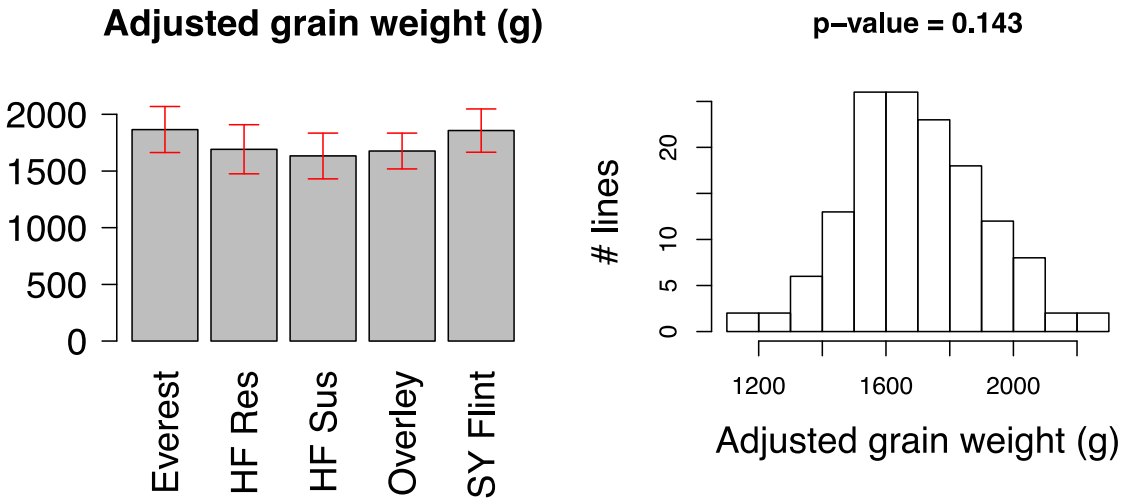

Figure 4.2. Distribution of mean adjusted grain weight in grams.

(Left panel) Barplot for two controls, 'Everest' and 'SY Flint', resistant lines, susceptible lines, and recurrent parent 'Overley'. (Right panel) Distribution of adjusted grain weight values for the whole population. Welch t-test *p*-value is shown on the top of histogram.

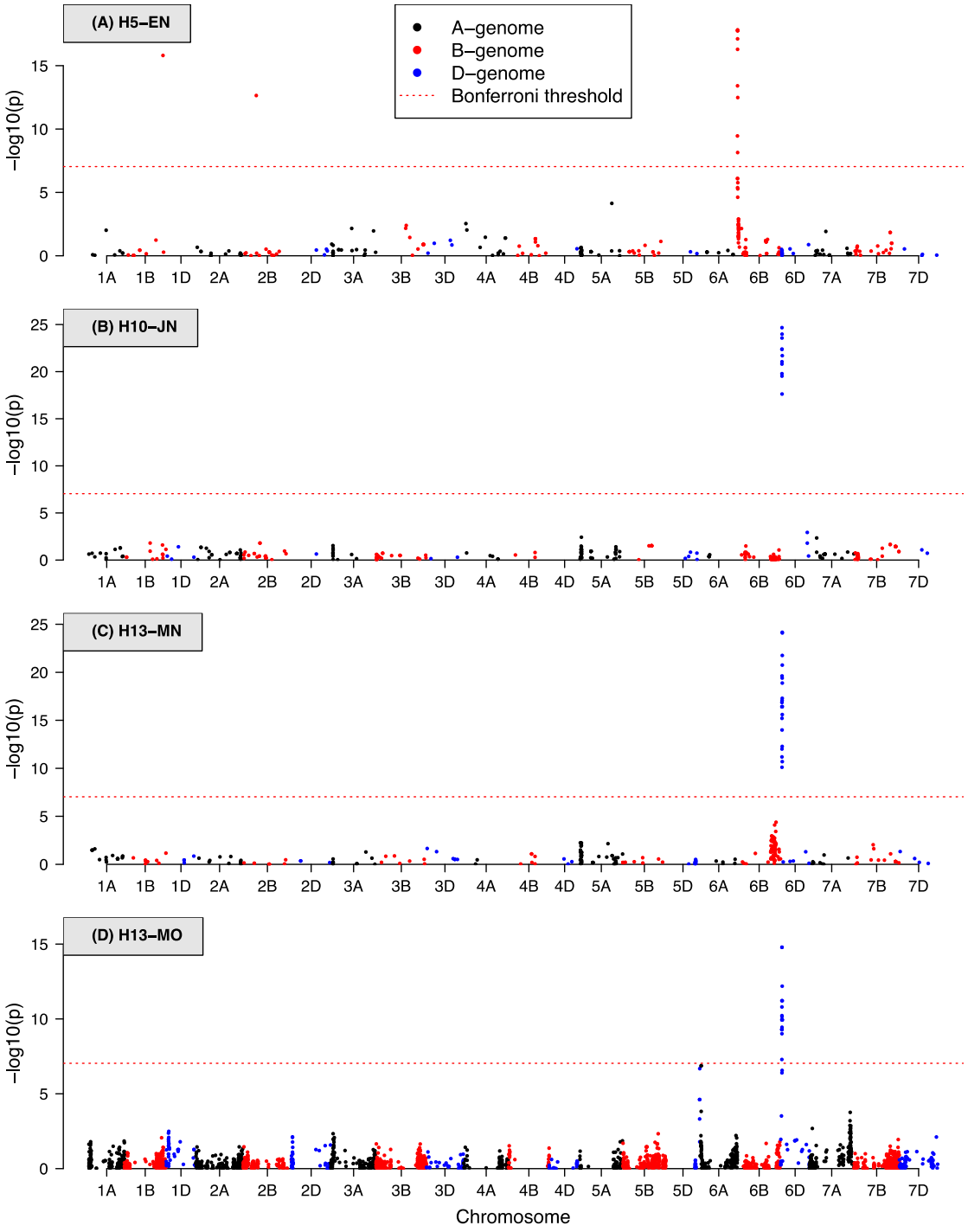

Figure 4.3. Manhattan plot showing the SNP associations for (A) H5-**EN, (B)** H10-**JN, (C)** H13-**MN, and (D)** H13-**MO.**

The 21 wheat chromosomes with physical positions are on the x-axis. Y-axis is the –log$_{10}$ of the *p*-value for each SNP. Red horizontal line is the Bonferroni threshold. Chromosome labels are placed in the middle of each chromosome. Chromosome labels are missing if no SNPs were detected on that chromosome.

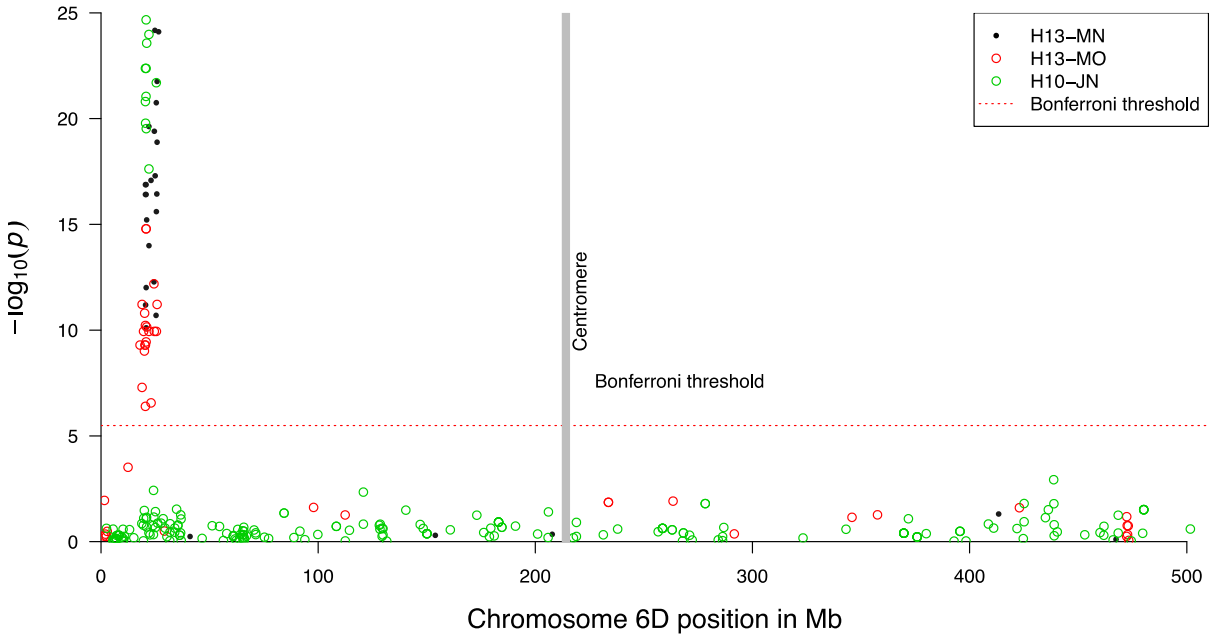

Figure 4.4. Joint Manhattan plot for *H10* and *H13* on chromosome 6D.

Grey vertical bar represents centromere. Red horizontal line is the Bonferroni's threshold.

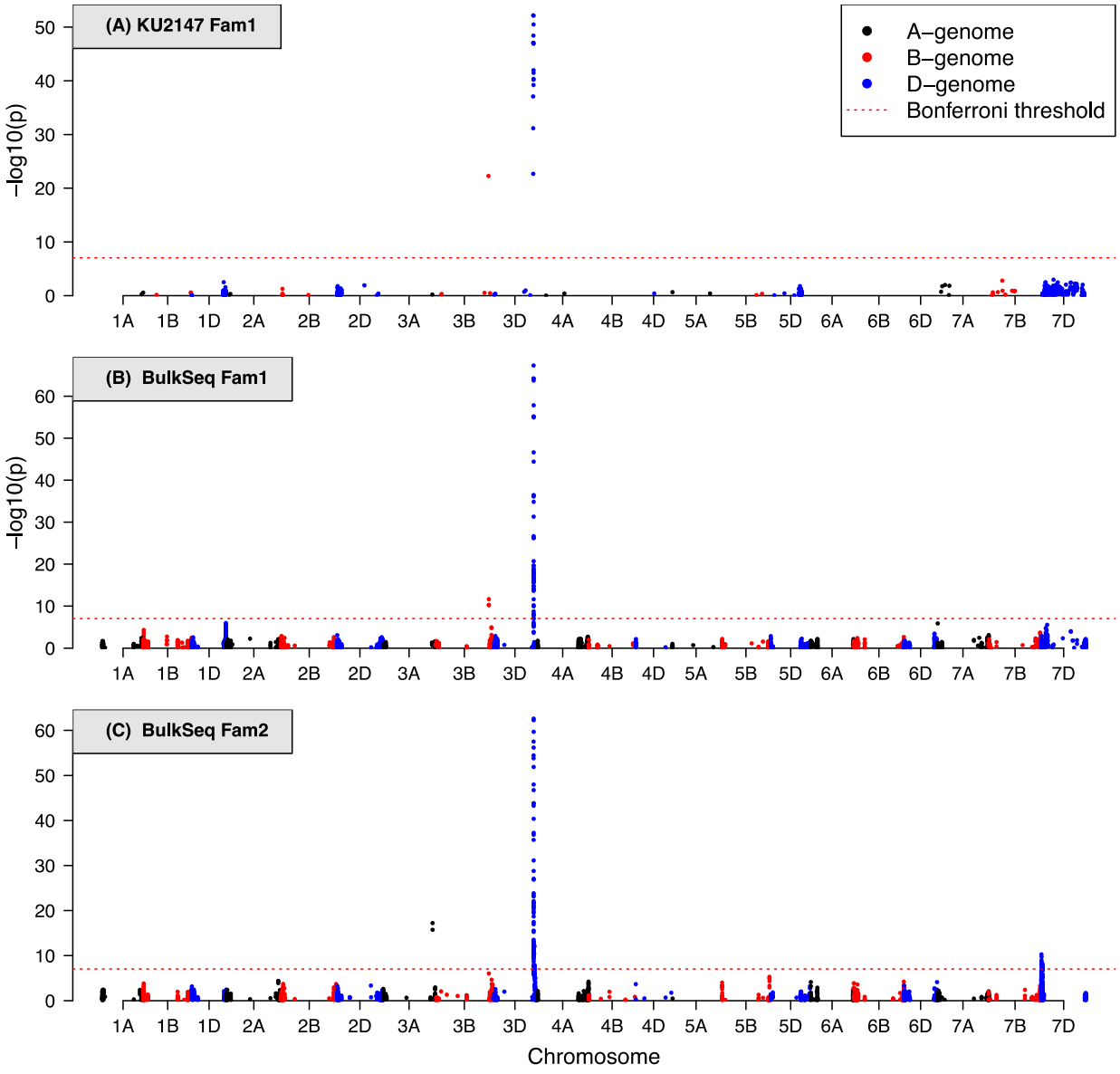

Figure 4.5. Manhattan plot showing the SNP distribution for newly introgressed gene from *Ae. tauschii* in (A) GBS *KU2147*, (B) RenSeq Fam1 and (C) RenSeq Fam2, along 21 wheat chromosomes on the x-axis.

Y-axis is the $-\log_{10}$ of the *p*-value for each SNP. Higher the SNP located on the y-axis, more is the association with HF resistance. Red horizontal line represents Bonferroni threshold level.

Appendix A - Copyright information

This appendix contains copyright information for licenses required to republish the content in this dissertation.

Figure 1.1 is taken from Ray *et al.* (2013) as it is, and we do not claim any rights on that image. Any requests for modifying and republishing should be directed to the original author.

Except Figure 1.1, all the content in this dissertation is the original research material that has not been published yet, and we reserve the right to publish it first. Permission is required from the original author before using, distributing and reproducing any content from this dissertation, and provided the work is properly credited.

Appendix B - Supplementary material Chapter 2

This appendix contains supplementary figures and tables for Chapter 2.

Table B.1. List of *Ae. tauschii* accessions from different gene banks

S.No.	*WGRC*	*CIMMYT*	*PAU*
1	TA10069	GID124083	PAUAT13
2	TA10070	GID155948	PAUAT13757
3	TA10071	GID156298	PAUAT13761
4	TA10072	GID156383	PAUAT13762
5	TA10073	GID156384	PAUAT13764
6	TA10074	GID156386	PAUAT13765
7	TA10075	GID156388	PAUAT13780
8	TA10076	GID156389	PAUAT13781
9	TA10077	GID156390	PAUAT14088
10	TA10078	GID156392	PAUAT14091
11	TA10079	GID156394	PAUAT14092
12	TA10080	GID156395	PAUAT14096
13	TA10081	GID156396	PAUAT14100
14	TA10082	GID156397	PAUAT14102
15	TA10083	GID156398	PAUAT14103
16	TA10084	GID156399	PAUAT14104
17	TA10085	GID156400	PAUAT14105
18	TA10086	GID156401	PAUAT14106
19	TA10087	GID156402	PAUAT14109
20	TA10088	GID156403	PAUAT14111
21	TA10089	GID156405	PAUAT14113
22	TA10090	GID156410	PAUAT14115
23	TA10091	GID156411	PAUAT14116
24	TA10092	GID156414	PAUAT14118
25	TA10093	GID156417	PAUAT14122
26	TA10094	GID156422	PAUAT14128
27	TA10095	GID156423	PAUAT14129
28	TA10096	GID156424	PAUAT14130
29	TA10097	GID156425	PAUAT14135
30	TA10098	GID156426	PAUAT14136
31	TA10099	GID156427	PAUAT14138
32	TA10100	GID156428	PAUAT14139
33	TA10101	GID156429	PAUAT14140
34	TA10102	GID156430	PAUAT14145
35	TA10103	GID156434	PAUAT14147
36	TA10104	GID156436	PAUAT14156
37	TA10105	GID156438	PAUAT14158
38	TA10106	GID156439	PAUAT14159
39	TA10107	GID156440	PAUAT14160
40	TA10108	GID156442	PAUAT14162
41	TA10109	GID156443	PAUAT14163
42	TA10110	GID156444	PAUAT14165
43	TA10112	GID156445	PAUAT14166
44	TA10113	GID156446	PAUAT14170
45	TA10114	GID156447	PAUAT14174
46	TA10115	GID156448	PAUAT14175
47	TA10116	GID156449	PAUAT14177
48	TA10117	GID156450	PAUAT14180
49	TA10118	GID156451	PAUAT14181
50	TA10119	GID156453	PAUAT14185
51	TA10120	GID156454	PAUAT14186
52	TA10121	GID156455	PAUAT14187
53	TA10122	GID156456	PAUAT14190
54	TA10123	GID156457	PAUAT14194
55	TA10124	GID156458	PAUAT14195
56	TA10125	GID156460	PAUAT14197
57	TA10126	GID156461	PAUAT14200
58	TA10127	GID156462	PAUAT14201
59	TA10128	GID156465	PAUAT14202
60	TA10129	GID156467	PAUAT14203
61	TA10130	GID156468	PAUAT14204
62	TA10131	GID156473	PAUAT14205
63	TA10132	GID156481	PAUAT14206
64	TA10133	GID156484	PAUAT14208
65	TA10134	GID156485	PAUAT14209
66	TA10135	GID156486	PAUAT14210
67	TA10136	GID156488	PAUAT14211
68	TA10137	GID156489	PAUAT14214
69	TA10138	GID156490	PAUAT14217
70	TA10139	GID156491	PAUAT14223
71	TA10140	GID156493	PAUAT14225
72	TA10141	GID156494	PAUAT14227
73	TA10142	GID156495	PAUAT14228
74	TA10143	GID156500	PAUAT14229
75	TA10144	GID156501	PAUAT14230
76	TA10145	GID156505	PAUAT14231
77	TA10148	GID156509	PAUAT14232
78	TA10155	GID156510	PAUAT14236
79	TA10156	GID156512	PAUAT14237
80	TA10157	GID156515	PAUAT14238
81	TA10158	GID156517	PAUAT14240
82	TA10159	GID156519	PAUAT14241
83	TA10160	GID156520	PAUAT14242

84	TA10161	GID156523	PAUAT14246	*135*	TA10296	GID156622	PAUAT14985
85	TA10162	GID156525	PAUAT14251	*136*	TA10297	GID156624	PAUAT14990
86	TA10163	GID156526	PAUAT14252	*137*	TA10298	GID156625	PAUAT14992
87	TA10164	GID156527	PAUAT14253	*138*	TA10299	GID156627	PAUAT14993
88	TA10165	GID156529	PAUAT14254	*139*	TA10300	GID156628	PAUAT14995
89	TA10166	GID156531	PAUAT14319	*140*	TA10301	GID156629	PAUAT14996
90	TA10167	GID156532	PAUAT14323	*141*	TA10302	GID156630	PAUAT14998
91	TA10168	GID156533	PAUAT14325	*142*	TA10303	GID156632	PAUAT14999
92	TA10169	GID156535	PAUAT14329	*143*	TA10304	GID156634	PAUAT17
93	TA10170	GID156536	PAUAT14330	*144*	TA10305	GID156635	PAUAT3544
94	TA10171	GID156537	PAUAT14334	*145*	TA10306	GID156636	PAUAT3733
95	TA10172	GID156539	PAUAT14336	*146*	TA10307	GID156637	PAUAT3735
96	TA10173	GID156542	PAUAT14337	*147*	TA10308	GID156641	PAUAT3742
97	TA10174	GID156543	PAUAT14338	*148*	TA10309	GID156642	PAUAT3744
98	TA10175	GID156546	PAUAT14339	*149*	TA10310	GID156643	PAUAT3746
99	TA10176	GID156549	PAUAT14340	*150*	TA10311	GID156644	PAUAT3750
100	TA10177	GID156550	PAUAT14341	*151*	TA10312	GID156646	PAUAT3751
101	TA10178	GID156552	PAUAT14343	*152*	TA10313	GID156647	PAUAT3752
102	TA10179	GID156556	PAUAT14345	*153*	TA10314	GID156651	PAUAT3753
103	TA10180	GID156557	PAUAT14347	*154*	TA10315	GID156652	PAUAT3755
104	TA10181	GID156559	PAUAT14348	*155*	TA10316	GID156654	PAUAT3757
105	TA10182	GID156562	PAUAT14351	*156*	TA10317	GID156655	PAUAT3758
106	TA10183	GID156572	PAUAT14352	*157*	TA10318	GID156656	PAUAT3759
107	TA10184	GID156574	PAUAT14353	*158*	TA10319	GID156658	PAUAT3760
108	TA10185	GID156577	PAUAT14354	*159*	TA10320	GID156666	PAUAT3761
109	TA10186	GID156578	PAUAT14355	*160*	TA10321	GID156668	PAUAT3766
110	TA10187	GID156579	PAUAT14356	*161*	TA10322	GID156669	PAUAT3769
111	TA10188	GID156582	PAUAT14359	*162*	TA10323	GID156671	PAUAT3784
112	TA10189	GID156583	PAUAT14360	*163*	TA10324	GID156672	PAUAT3798
113	TA10190	GID156584	PAUAT14362	*164*	TA10325	GID156675	PAUAT3799
114	TA10191	GID156586	PAUAT14576	*165*	TA10326	GID156676	PAUAT3805
115	TA10192	GID156587	PAUAT14578	*166*	TA10327	GID156677	PAUAT3806
116	TA10193	GID156588	PAUAT14582	*167*	TA10328	GID156678	PAUAT3822
117	TA10194	GID156589	PAUAT14583	*168*	TA10329	GID156679	PAUAT3823
118	TA10195	GID156590	PAUAT14586	*169*	TA10330	GID156681	PAUAT3826
119	TA10196	GID156591	PAUAT14953	*170*	TA10331	GID156682	PAUAT9785
120	TA10197	GID156592	PAUAT14954	*171*	TA10417	GID156683	PAUAT9787
121	TA10198	GID156594	PAUAT14957	*172*	TA10836	GID156684	PAUAT9788
122	TA10199	GID156595	PAUAT14958	*173*	TA10837	GID156686	PAUAT9790
123	TA10200	GID156596	PAUAT14960	*174*	TA10838	GID156689	PAUAT9791
124	TA10202	GID156597	PAUAT14962	*175*	TA10839	GID156690	PAUAT9795
125	TA10205	GID156600	PAUAT14966	*176*	TA10869	GID156691	PAUAT9796
126	TA10210	GID156609	PAUAT14967	*177*	TA10871	GID156694	PAUAT9798
127	TA10211	GID156611	PAUAT14968	*178*	TA10872	GID156695	PAUAT9800
128	TA10212	GID156612	PAUAT14970	*179*	TA10918	GID156698	PAUAT9804
129	TA10213	GID156613	PAUAT14971	*180*	TA10919	GID156699	PAUAT9806
130	TA10291	GID156614	PAUAT14972	*181*	TA10920	GID156701	PAUAT9807
131	TA10292	GID156615	PAUAT14973	*182*	TA10921	GID156702	PAUAT9809
132	TA10293	GID156618	PAUAT14974	*183*	TA10922	GID156703	PAUAT9822
133	TA10294	GID156619	PAUAT14975	*184*	TA10923	GID156705	PAUAT9823
134	TA10295	GID156620	PAUAT14979	*185*	TA10924	GID156707	PAUAT9824

186	TA10925	GID156708	PAUAT9829
187	TA10926	GID156709	PAUAT9830
188	TA10927	GID156711	
189	TA10928	GID156712	
190	TA10929	GID156713	
191	TA10930	GID156714	
192	TA10931	GID156715	
193	TA10932	GID156718	
194	TA10933	GID156719	
195	TA10934	GID156720	
196	TA10935	GID156721	
197	TA10936	GID156722	
198	TA10937	GID156724	
199	TA10938	GID156725	
200	TA10939	GID156726	
201	TA10940	GID156727	
202	TA10941	GID156729	
203	TA10942	GID156730	
204	TA10943	GID156731	
205	TA10944	GID156733	
206	TA10945	GID156734	
207	TA10946	GID156735	
208	TA10947	GID156736	
209	TA10948	GID156738	
210	TA10949	GID156739	
211	TA10950	GID156740	
212	TA10951	GID156741	
213	TA10952	GID156742	
214	TA10953	GID156743	
215	TA10954	GID156750	
216	TA10955	GID156754	
217	TA10956	GID156764	
218	TA10957	GID156766	
219	TA10958	GID156767	
220	TA10959	GID156769	
221	TA10960	GID156778	
222	TA10961	GID156780	
223	TA11020	GID156782	
224	TA11021	GID156790	
225	TA1577	GID156791	
226	TA1578	GID156792	
227	TA1579	GID156804	
228	TA1580	GID156805	
229	TA1581	GID156811	
230	TA1582	GID156814	
231	TA1583	GID156823	
232	TA1584	GID156825	
233	TA1585	GID156827	
234	TA1586	GID156830	
235	TA1587	GID156835	
236	TA1588	GID156849	
237	TA1589	GID156850	
238	TA1590	GID156854	
239	TA1591	GID156868	
240	TA1592	GID156871	
241	TA1593	GID156879	
242	TA1594	GID156886	
243	TA1595	GID156894	
244	TA1596	GID156904	
245	TA1597	GID156905	
246	TA1598	GID156906	
247	TA1599	GID156975	
248	TA1600	GID156977	
249	TA1601	GID156978	
250	TA1602	GID156979	
251	TA1603	GID156980	
252	TA1604	GID156981	
253	TA1605	GID156982	
254	TA1606	GID156983	
255	TA1612	GID156984	
256	TA1613	GID156985	
257	TA1615	GID156986	
258	TA1616	GID156987	
259	TA1617	GID156988	
260	TA1618	GID156989	
261	TA1619	GID156990	
262	TA1620	GID156991	
263	TA1621	GID156992	
264	TA1622	GID156993	
265	TA1623	GID156995	
266	TA1624	GID156996	
267	TA1625	GID156997	
268	TA1626	GID156998	
269	TA1629	GID157000	
270	TA1630	GID157001	
271	TA1631	GID157002	
272	TA1632	GID157003	
273	TA1633	GID157004	
274	TA1634	GID157005	
275	TA1635	GID157006	
276	TA1639	GID157007	
277	TA1640	GID157008	
278	TA1641	GID157009	
279	TA1642	GID157010	
280	TA1643	GID157011	
281	TA1644	GID157013	
282	TA1645	GID157014	
283	TA1649	GID157016	
284	TA1650	GID157017	
285	TA1651	GID157018	
286	TA1652	GID157019	
287	TA1653	GID157020	

288	TA1655	GID157021
289	TA1656	GID157022
290	TA1657	GID157023
291	TA1658	GID157024
292	TA1659	GID157025
293	TA1660	GID157026
294	TA1661	GID157027
295	TA1662	GID157028
296	TA1664	GID157029
297	TA1665	GID157030
298	TA1666	GID157032
299	TA1667	GID157033
300	TA1668	GID157034
301	TA1669	GID157035
302	TA1670	GID157036
303	TA1671	GID157037
304	TA1672	GID157038
305	TA1673	GID157039
306	TA1674	GID157040
307	TA1675	GID157041
308	TA1676	GID157042
309	TA1677	GID157043
310	TA1678	GID157044
311	TA1679	GID157045
312	TA1680	GID157047
313	TA1681	GID157051
314	TA1682	GID157053
315	TA1683	GID157054
316	TA1684	GID157056
317	TA1685	GID157057
318	TA1686	GID157058
319	TA1687	GID157059
320	TA1688	GID157062
321	TA1689	GID157064
322	TA1690	GID157065
323	TA1691	GID157067
324	TA1692	GID157068
325	TA1693	GID157069
326	TA1694	GID157072
327	TA1695	GID157074
328	TA1696	GID157077
329	TA1697	GID157078
330	TA1698	GID157079
331	TA1699	GID157080
332	TA1700	GID157081
333	TA1703	GID157082
334	TA1704	GID157083
335	TA1706	GID157085
336	TA1707	GID157086
337	TA1708	GID157087
338	TA1709	GID157088
339	TA1712	GID157090
340	TA1713	GID157092
341	TA1714	GID157093
342	TA1715	GID157094
343	TA1716	GID157095
344	TA1717	GID157096
345	TA1718	GID157097
346	TA2118	GID157098
347	TA2119	GID157099
348	TA2123	GID157102
349	TA2369	GID157103
350	TA2370	GID157104
351	TA2371	GID157105
352	TA2372	GID157108
353	TA2373	GID157109
354	TA2374	GID157110
355	TA2375	GID157111
356	TA2376	GID157112
357	TA2377	GID157114
358	TA2378	GID157116
359	TA2379	GID157117
360	TA2380	GID157122
361	TA2381	GID157124
362	TA2382	GID157125
363	TA2383	GID157127
364	TA2384	GID157130
365	TA2385	GID157132
366	TA2386	GID157134
367	TA2387	GID157135
368	TA2388	GID157136
369	TA2389	GID157137
370	TA2390	GID157138
371	TA2391	GID157140
372	TA2392	GID157141
373	TA2393	GID157142
374	TA2394	GID157143
375	TA2395	GID157145
376	TA2396	GID157146
377	TA2397	GID157148
378	TA2398	GID157149
379	TA2399	GID157150
380	TA2400	GID157151
381	TA2401	GID157425
382	TA2402	GID68345
383	TA2403	GID68347
384	TA2404	GID68357
385	TA2405	GID68375
386	TA2406	GID69369
387	TA2407	GID95324
388	TA2408	GID95325
389	TA2409	

390	TA2410
391	TA2411
392	TA2412
393	TA2413
394	TA2414
395	TA2415
396	TA2416
397	TA2417
398	TA2418
399	TA2419
400	TA2420
401	TA2421
402	TA2422
403	TA2423
404	TA2424
405	TA2425
406	TA2426
407	TA2427
408	TA2428
409	TA2429
410	TA2430
411	TA2431
412	TA2432
413	TA2433
414	TA2434
415	TA2435
416	TA2436
417	TA2437
418	TA2438
419	TA2439
420	TA2440
421	TA2441
422	TA2442
423	TA2443
424	TA2444
425	TA2445
426	TA2446
427	TA2447
428	TA2448
429	TA2449
430	TA2450
431	TA2451
432	TA2452
433	TA2453
434	TA2454
435	TA2455
436	TA2456
437	TA2457
438	TA2458
439	TA2459
440	TA2460
441	TA2461
442	TA2462
443	TA2463
444	TA2464
445	TA2465
446	TA2466
447	TA2467
448	TA2468
449	TA2469
450	TA2470
451	TA2471
452	TA2472
453	TA2473
454	TA2474
455	TA2475
456	TA2476
457	TA2477
458	TA2478
459	TA2479
460	TA2480
461	TA2481
462	TA2482
463	TA2483
464	TA2484
465	TA2485
466	TA2486
467	TA2487
468	TA2488
469	TA2489
470	TA2490
471	TA2491
472	TA2492
473	TA2493
474	TA2494
475	TA2495
476	TA2496
477	TA2497
478	TA2498
479	TA2499
480	TA2500
481	TA2501
482	TA2502
483	TA2503
484	TA2504
485	TA2505
486	TA2506
487	TA2507
488	TA2508
489	TA2509
490	TA2510
491	TA2511

492	TA2512
493	TA2513
494	TA2514
495	TA2515
496	TA2516
497	TA2517
498	TA2518
499	TA2519
500	TA2520
501	TA2521
502	TA2522
503	TA2523
504	TA2524
505	TA2525
506	TA2526
507	TA2527
508	TA2528
509	TA2529
510	TA2530
511	TA2531
512	TA2532
513	TA2533
514	TA2534
515	TA2535
516	TA2536
517	TA2537
518	TA2538
519	TA2539
520	TA2540
521	TA2541
522	TA2542
523	TA2543
524	TA2544
525	TA2545
526	TA2546
527	TA2547
528	TA2548
529	TA2549
530	TA2550
531	TA2551
532	TA2552
533	TA2553
534	TA2554
535	TA2555
536	TA2556
537	TA2557
538	TA2558
539	TA2559
540	TA2560
541	TA2561
542	TA2562
543	TA2563
544	TA2564
545	TA2565
546	TA2566
547	TA2567
548	TA2568
549	TA2569
550	TA2570
551	TA2571
552	TA2572
553	TA2573
554	TA2574
555	TA2575
556	TA2576
557	TA2577
558	TA2578
559	TA2579
560	TA2580
561	TA2581
562	TA2582
563	TA2583
564	TA2584
565	TA2585
566	TA2586
567	TA2587
568	TA3429

Table B.2. List of matching *Ae. tauschii* accessions.

Legend: Grp# = arbitrary group number; Grp size = # accessions in group; columns 3-5 (W=WGRC, P=PAU, C=CIMMYT) = # accessions in a group from individual genebanks; Matching accessions = accessions matching 99% or more within a group.

Grp#	*Grp size*	*W*	*P*	*C*	*Matching accessions*
Grp1	4	1	1	2	GID124083, GID156584, PAUAT14217, TA2530
Grp2	2	0	0	2	GID155948, GID156386
Grp3	6	0	1	5	GID156298, GID156383, GID156389, GID156398, GID156741, PAUAT14325
Grp4	2	0	0	2	GID156384, GID156400
Grp5	5	2	0	3	GID156388, GID156461, GID156734, TA10078, TA2386
Grp6	17	3	3	11	GID156390, GID156533, GID156594, GID156619, GID156719, GID156750, GID156754, GID156987, GID157017, GID68345, GID68347, PAUAT13, PAUAT14957, PAUAT17, TA1600, TA2463, TA2464
Grp7	2	0	0	2	GID156392, GID156402
Grp8	1	0	0	1	GID156394
Grp9	3	0	2	1	GID156395, PAUAT13780, PAUAT13781
Grp10	1	0	0	1	GID156396
Grp11	1	0	0	1	GID156397
Grp12	1	0	0	1	GID156399
Grp13	1	0	0	1	GID156401
Grp14	1	0	0	1	GID156403
Grp15	8	3	1	4	GID156405, GID156611, GID156718, GID157141, PAUAT14995, TA1635, TA1695, TA2457
Grp16	8	3	1	4	GID156410, GID156671, GID156736, GID157082, PAUAT14103, TA10117, TA1598, TA1657
Grp17	9	6	0	3	GID156411, GID157116, GID157127, TA10101, TA1653, TA1665, TA1667, TA1678, TA1687
Grp18	1	0	0	1	GID156414
Grp19	5	1	1	3	GID156417, GID156625, GID157088, PAUAT14975, TA1623
Grp20	17	7	6	4	GID156422, GID156489, GID156494, GID156814, PAUAT13757, PAUAT13765, PAUAT14251, PAUAT14254, PAUAT9824, PAUAT9830, TA10139, TA10141, TA1604, TA2416, TA2421, TA2423, TA2424
Grp21	27	3	3	21	GID156423, GID156439, GID156440, GID156442, GID156444, GID156449, GID156460, GID156517, GID156595, GID156609, GID156637, GID156654, GID156666, GID156669, GID156735, GID156740, GID156767, GID156792, GID156849, GID156868, GID157053, PAUAT14953, PAUAT3733, PAUAT3735, TA2118, TA2385, TA2468
Grp22	9	1	0	8	GID156424, GID156425, GID156448, GID156468, GID156708, GID156979, GID156982, GID157143, TA2394
Grp23	10	5	0	5	GID156426, GID156509, GID156711, GID156712, GID156981, TA2422, TA2431, TA2435, TA2436, TA2437

Grp24	10	2	5	3	GID156427, GID156428, GID95324, PAUAT14954, PAUAT14973, PAUAT14974, PAUAT14992, PAUAT3746, TA1651, TA2449
Grp25	3	0	0	3	GID156429, GID156634, GID157016
Grp26	10	3	0	7	GID156430, GID156490, GID156525, GID156526, GID156527, GID156713, GID157104, TA2452, TA2454, TA2455
Grp27	1	0	0	1	GID156434
Grp28	13	1	8	4	GID156436, GID156535, GID156647, GID156805, PAUAT14334, PAUAT14343, PAUAT14347, PAUAT14958, PAUAT3758, PAUAT9804, PAUAT9806, PAUAT9807, TA2465
Grp29	15	2	8	5	GID156438, GID156531, GID156543, GID156886, GID156992, PAUAT14115, PAUAT14139, PAUAT14145, PAUAT14323, PAUAT14359, PAUAT14576, PAUAT14960, PAUAT9823, TA1581, TA2474
Grp30	6	2	3	1	GID156443, PAUAT14962, PAUAT14999, PAUAT3766, TA2486, TA2487
Grp31	8	3	2	3	GID156445, GID156446, GID156998, PAUAT14966, PAUAT14985, TA10133, TA2370, TA2372
Grp32	2	1	0	1	GID156447, TA2371
Grp33	11	1	9	1	GID156450, PAUAT14214, PAUAT14351, PAUAT14352, PAUAT14353, PAUAT14354, PAUAT14578, PAUAT3753, PAUAT3759, PAUAT3826, TA2374
Grp34	5	3	0	2	GID156451, GID156519, TA1690, TA2375, TA2446
Grp35	7	1	2	4	GID156453, GID156701, GID156702, GID156703, PAUAT14967, PAUAT14998, TA2377
Grp36	4	2	0	2	GID156454, GID156991, TA1599, TA2378
Grp37	18	5	6	7	GID156455, GID156456, GID156652, GID156977, GID156978, GID157134, GID157135, PAUAT14177, PAUAT14197, PAUAT3742, PAUAT9791, PAUAT9795, PAUAT9822, TA1688, TA1689, TA2379, TA2380, TA2388
Grp38	2	1	0	1	GID156457, TA2381
Grp39	2	1	0	1	GID156458, TA2382
Grp40	2	1	0	1	GID156462, TA2387
Grp41	4	2	0	2	GID156465, GID157054, TA2390, TA2391
Grp42	6	1	3	2	GID156467, GID157138, PAUAT14104, PAUAT14105, PAUAT14106, TA1692
Grp43	14	2	9	3	GID156473, GID156993, GID157019, PAUAT14122, PAUAT14136, PAUAT14165, PAUAT14210, PAUAT14211, PAUAT14227, PAUAT14341, PAUAT14586, PAUAT3784, TA1703, TA2475
Grp44	3	1	0	2	GID156481, GID156577, TA2520
Grp45	4	2	0	2	GID156484, GID156485, TA2411, TA2412
Grp46	2	1	0	1	GID156486, TA2413
Grp47	4	2	0	2	GID156488, GID157085, TA1620, TA2415
Grp48	1	0	0	1	GID156491
Grp49	2	1	0	1	GID156493, TA2420
Grp50	2	1	0	1	GID156495, TA2425
Grp51	3	1	0	2	GID156500, GID156501, TA2427
Grp52	2	1	0	1	GID156505, TA2431

Grp53	1	0	0	1	GID156510
Grp54	2	1	0	1	GID156512, TA2439
Grp55	3	1	0	2	GID156515, GID156536, TA2442
Grp56	3	2	0	1	GID156520, TA2447, TA2448
Grp57	4	2	0	2	GID156523, GID156658, TA2450, TA2451
Grp58	5	2	0	3	GID156529, GID156714, GID156984, TA1605, TA2456
Grp59	2	1	0	1	GID156532, TA2494
Grp60	4	2	0	2	GID156537, GID157018, TA2467, TA2469
Grp61	44	3	31	10	GID156539, GID156559, GID156613, GID156622, GID156721, GID156722, GID156988, GID156989, GID156990, GID68357, PAUAT13761, PAUAT14129, PAUAT14135, PAUAT14170, PAUAT14181, PAUAT14185, PAUAT14186, PAUAT14190, PAUAT14223, PAUAT14232, PAUAT14241, PAUAT14246, PAUAT14319, PAUAT14329, PAUAT14330, PAUAT14336, PAUAT14337, PAUAT14338, PAUAT14339, PAUAT14340, PAUAT14356, PAUAT14362, PAUAT14582, PAUAT14583, PAUAT3757, PAUAT3760, PAUAT3761, PAUAT3769, PAUAT3798, PAUAT3799, PAUAT3805, TA2468, TA2470, TA2472
Grp62	2	1	0	1	GID156542, TA2473
Grp63	2	1	0	1	GID156546, TA2477
Grp64	6	2	0	4	GID156549, GID156550, GID157024, GID157096, TA2476, TA2481
Grp65	3	1	0	2	GID156552, GID156572, TA2536
Grp66	2	1	0	1	GID156556, TA2488
Grp67	3	2	0	1	GID156557, TA2490, TA2491
Grp68	3	1	0	2	GID156562, GID157021, TA2495
Grp69	3	1	0	2	GID156574, GID157022, TA2496
Grp70	2	1	0	1	GID156578, TA2523
Grp71	1	0	0	1	GID156579
Grp72	1	0	0	1	GID156582
Grp73	2	1	0	1	GID156583, TA2528
Grp74	3	1	0	2	GID156586, GID156743, TA2539
Grp75	2	1	0	1	GID156587, TA2540
Grp76	2	1	0	1	GID156588, TA2541
Grp77	2	1	0	1	GID156589, TA2543
Grp78	2	1	0	1	GID156590, TA2549
Grp79	1	0	0	1	GID156591
Grp80	2	1	0	1	GID156592, TA2554
Grp81	1	0	0	1	GID156596
Grp82	2	1	0	1	GID156597, TA2563
Grp83	7	1	0	6	GID156600, GID156620, GID156975, GID157030, GID157059, GID157142, TA1696
Grp84	6	2	1	3	GID156612, GID156635, GID156983, PAUAT14209, TA1644, TA2453
Grp85	2	0	0	2	GID156614, GID156615
Grp86	1	0	0	1	GID156618
Grp87	3	1	0	2	GID156624, GID157087, TA1622
Grp88	1	0	0	1	GID156627

Grp89	1	0	0	1	GID156628
Grp90	2	1	0	1	GID156629, TA2406
Grp91	2	1	0	1	GID156630, TA2407
Grp92	1	0	0	1	GID156632
Grp93	4	2	0	2	GID156636, GID156985, TA2458, TA2462
Grp94	3	1	0	2	GID156641, GID157020, TA2478
Grp95	1	0	0	1	GID156642
Grp96	5	1	0	4	GID156643, GID156804, GID156811, GID157099, TA1645
Grp97	1	0	0	1	GID156644
Grp98	3	2	0	1	GID156651, TA2428, TA2434
Grp99	2	1	0	1	GID156655, TA2397
Grp100	2	1	0	1	GID156656, TA1613
Grp101	3	1	1	1	GID156668, PAUAT14163, TA1617
Grp102	1	0	0	1	GID156672
Grp103	1	0	0	1	GID156675
Grp104	2	0	0	2	GID156676, GID157006
Grp105	2	0	0	2	GID156677, GID156678
Grp106	1	0	0	1	GID156679
Grp107	1	0	0	1	GID156681
Grp108	1	0	0	1	GID156682
Grp109	1	0	0	1	GID156683
Grp110	1	0	0	1	GID156684
Grp111	2	1	0	1	GID156686, TA10071
Grp112	2	1	0	1	GID156689, TA10075
Grp113	3	1	1	1	GID156690, PAUAT14111, TA1639
Grp114	1	0	0	1	GID156691
Grp115	1	0	0	1	GID156694
Grp116	1	0	0	1	GID156695
Grp117	1	0	0	1	GID156698
Grp118	1	0	0	1	GID156699
Grp119	2	1	0	1	GID156707, TA2392
Grp120	5	1	0	4	GID156709, GID156871, GID156980, GID157056, TA2402
Grp121	1	0	0	1	GID156715
Grp122	1	0	0	1	GID156720
Grp123	1	0	0	1	GID156725
Grp124	1	0	0	1	GID156726
Grp125	1	0	0	1	GID156727
Grp126	9	3	0	6	GID156729, GID156780, GID156879, GID157079, GID157081, GID157095, TA1595, TA1597, TA1634
Grp127	1	0	0	1	GID156730
Grp128	1	0	0	1	GID156731
Grp129	1	0	0	1	GID156733
Grp130	3	0	1	2	GID156738, GID156739, PAUAT9829
Grp131	3	1	0	2	GID156742, GID156904, TA2537
Grp132	7	5	0	2	GID156764, GID157146, TA10100, TA1601, TA1700, TA2567, TA2569
Grp133	10	8	0	2	GID156766, GID157146, TA10100, TA10129, TA1601, TA1700, TA2567, TA2569, TA2570, TA2575
Grp134	2	1	0	1	GID156769, TA2584
Grp135	1	0	0	1	GID156778

Grp136	7	1	4	2	GID156782, GID157114, PAUAT14128, PAUAT3544, PAUAT9788, PAUAT9790, TA1664
Grp137	2	1	0	1	GID156790, TA1626
Grp138	1	0	0	1	GID156791
Grp139	1	0	0	1	GID156823
Grp140	1	0	0	1	GID156825
Grp141	3	1	0	2	GID156827, GID156850, TA10081
Grp142	2	1	0	1	GID156830, TA10123
Grp143	2	0	0	2	GID156835, GID69369
Grp144	2	1	0	1	GID156854, TA1717
Grp145	1	0	0	1	GID156894
Grp146	1	0	0	1	GID156905
Grp147	7	3	3	1	GID156906, PAUAT14174, PAUAT3744, PAUAT3806, TA10140, TA1602, TA1603
Grp148	1	0	0	1	GID156995
Grp149	2	1	0	1	GID156996, TA2369
Grp150	1	0	0	1	GID156997
Grp151	1	0	0	1	GID157000
Grp152	1	0	0	1	GID157001
Grp153	1	0	0	1	GID157002
Grp154	1	0	0	1	GID157003
Grp155	1	0	0	1	GID157004
Grp156	1	0	0	1	GID157005
Grp157	1	0	0	1	GID157007
Grp158	1	0	0	1	GID157008
Grp159	1	0	0	1	GID157009
Grp160	1	0	0	1	GID157010
Grp161	1	0	0	1	GID157011
Grp162	1	0	0	1	GID157013
Grp163	1	0	0	1	GID157014
Grp164	2	1	0	1	GID157023, TA2499
Grp165	3	1	0	2	GID157025, GID157097, TA2482
Grp166	2	1	0	1	GID157026, TA2445
Grp167	1	0	0	1	GID157027
Grp168	3	2	0	1	GID157028, TA1680, TA1686
Grp169	1	0	0	1	GID157029
Grp170	3	1	0	2	GID157032, GID157109, TA1658
Grp171	1	0	0	1	GID157033
Grp172	2	1	0	1	GID157034, TA10106
Grp173	1	0	0	1	GID157035
Grp174	1	0	0	1	GID157036
Grp175	1	0	0	1	GID157037
Grp176	2	1	0	1	GID157038, TA2521
Grp177	2	0	0	2	GID157039, GID157040
Grp178	1	0	0	1	GID157041
Grp179	1	0	0	1	GID157042
Grp180	1	0	0	1	GID157043
Grp181	3	1	0	2	GID157044, GID157045, TA2522
Grp182	1	0	0	1	GID157047
Grp183	1	0	0	1	GID157051
Grp184	2	1	0	1	GID157057, TA2408

Grp185	1	0	0	1	GID157058
Grp186	4	2	0	2	GID157062, GID157150, TA1577, TA1707
Grp187	4	2	0	2	GID157064, GID157092, TA1579, TA1629
Grp188	5	4	0	1	GID157065, TA1578, TA1580, TA1581, TA1655
Grp189	7	4	0	3	GID157067, GID157068, GID157069, TA1582, TA1583, TA1584, TA1585
Grp190	9	8	0	1	GID157072, TA1588, TA1589, TA2505, TA2506, TA2507, TA2508, TA2509, TA2510
Grp191	3	2	0	1	GID157074, TA1590, TA1591
Grp192	2	1	0	1	GID157077, TA1594
Grp193	2	1	0	1	GID157078, TA1593
Grp194	2	1	0	1	GID157080, TA1596
Grp195	2	1	0	1	GID157083, TA1618
Grp196	1	0	0	1	GID157086
Grp197	3	2	0	1	GID157090, TA1621, TA1625
Grp198	2	1	0	1	GID157093, TA1631
Grp199	2	1	0	1	GID157094, TA1632
Grp200	2	1	0	1	GID157098, TA2527
Grp201	2	1	0	1	GID157102, TA1649
Grp202	7	3	2	2	GID157103, GID157151, PAUAT14100, PAUAT14225, TA1650, TA1708, TA1709
Grp203	1	0	0	1	GID157105
Grp204	1	0	0	1	GID157108
Grp205	5	4	0	1	GID157110, TA10124, TA1615, TA1624, TA1659
Grp206	2	1	0	1	GID157111, TA1660
Grp207	6	1	4	1	GID157112, PAUAT14113, PAUAT14159, PAUAT14160, PAUAT9787, TA1661
Grp208	1	0	0	1	GID157117
Grp209	1	0	0	1	GID157122
Grp210	3	2	0	1	GID157124, TA1675, TA1693
Grp211	2	1	0	1	GID157125, TA1676
Grp212	2	1	0	1	GID157130, TA1681
Grp213	2	1	0	1	GID157132, TA1686
Grp214	1	0	0	1	GID157136
Grp215	3	1	0	2	GID157137, GID157148, TA1704
Grp216	2	1	0	1	GID157140, TA1694
Grp217	2	1	0	1	GID157145, TA1699
Grp218	3	1	1	1	GID157149, PAUAT14996, TA1706
Grp219	1	0	0	1	GID157425
Grp220	3	2	0	1	GID68375, TA2534, TA2535
Grp221	3	2	0	1	GID95325, TA10069, TA2532
Grp222	1	0	1	0	PAUAT13764
Grp223	1	0	1	0	PAUAT14088
Grp224	1	0	1	0	PAUAT14091
Grp225	1	0	1	0	PAUAT14092
Grp226	1	0	1	0	PAUAT14096
Grp227	1	0	1	0	PAUAT14102
Grp228	1	0	1	0	PAUAT14109
Grp229	1	0	1	0	PAUAT14116
Grp230	1	0	1	0	PAUAT14118
Grp231	1	0	1	0	PAUAT14130

Grp232	1	0	1	0	PAUAT14138
Grp233	4	1	3	0	PAUAT14140, PAUAT14337, PAUAT14583, TA2468
Grp234	2	0	2	0	PAUAT14147, PAUAT14238
Grp235	1	0	1	0	PAUAT14156
Grp236	1	0	1	0	PAUAT14158
Grp237	1	0	1	0	PAUAT14162
Grp238	1	0	1	0	PAUAT14166
Grp239	1	0	1	0	PAUAT14175
Grp240	2	0	2	0	PAUAT14180, PAUAT3751
Grp241	1	0	1	0	PAUAT14187
Grp242	1	0	1	0	PAUAT14194
Grp243	1	0	1	0	PAUAT14195
Grp244	1	0	1	0	PAUAT14200
Grp245	1	0	1	0	PAUAT14201
Grp246	1	0	1	0	PAUAT14202
Grp247	1	0	1	0	PAUAT14203
Grp248	3	1	2	0	PAUAT14204, PAUAT3823, TA1642
Grp249	1	0	1	0	PAUAT14205
Grp250	1	0	1	0	PAUAT14206
Grp251	1	0	1	0	PAUAT14208
Grp252	4	2	2	0	PAUAT14228, PAUAT14229, TA10181, TA10205
Grp253	2	1	1	0	PAUAT14230, TA10108
Grp254	1	0	1	0	PAUAT14231
Grp255	1	0	1	0	PAUAT14236
Grp256	1	0	1	0	PAUAT14237
Grp257	1	0	1	0	PAUAT14240
Grp258	1	0	1	0	PAUAT14242
Grp259	1	0	1	0	PAUAT14252
Grp260	1	0	1	0	PAUAT14253
Grp261	1	0	1	0	PAUAT14348
Grp262	2	0	2	0	PAUAT14355, PAUAT9785
Grp263	1	0	1	0	PAUAT14360
Grp264	1	0	1	0	PAUAT14968
Grp265	2	1	1	0	PAUAT14970, TA1632
Grp266	1	0	1	0	PAUAT14971
Grp267	1	0	1	0	PAUAT14972
Grp268	2	1	1	0	PAUAT14979, TA1587
Grp269	1	0	1	0	PAUAT14990
Grp270	1	0	1	0	PAUAT14993
Grp271	1	0	1	0	PAUAT3750
Grp272	1	0	1	0	PAUAT3752
Grp273	1	0	1	0	PAUAT3755
Grp274	1	0	1	0	PAUAT3822
Grp275	1	0	1	0	PAUAT9796
Grp276	1	0	1	0	PAUAT9798
Grp277	1	0	1	0	PAUAT9800
Grp278	1	0	1	0	PAUAT9809
Grp279	1	1	0	0	TA10070
Grp280	2	2	0	0	TA10072, TA2123
Grp281	1	1	0	0	TA10073
Grp282	1	1	0	0	TA10074

Grp283	1	1	0	0	TA10076
Grp284	1	1	0	0	TA10077
Grp285	1	1	0	0	TA10079
Grp286	1	1	0	0	TA10080
Grp287	1	1	0	0	TA10082
Grp288	1	1	0	0	TA10083
Grp289	1	1	0	0	TA10084
Grp290	2	2	0	0	TA10085, TA10086
Grp291	1	1	0	0	TA10087
Grp292	1	1	0	0	TA10088
Grp293	1	1	0	0	TA10089
Grp294	1	1	0	0	TA10090
Grp295	1	1	0	0	TA10091
Grp296	1	1	0	0	TA10092
Grp297	1	1	0	0	TA10093
Grp298	1	1	0	0	TA10094
Grp299	3	3	0	0	TA10095, TA10130, TA10131
Grp300	1	1	0	0	TA10096
Grp301	1	1	0	0	TA10097
Grp302	1	1	0	0	TA10098
Grp303	1	1	0	0	TA10099
Grp304	1	1	0	0	TA10102
Grp305	2	2	0	0	TA10103, TA2580
Grp306	1	1	0	0	TA10104
Grp307	1	1	0	0	TA10105
Grp308	1	1	0	0	TA10107
Grp309	2	2	0	0	TA10109, TA1623
Grp310	1	1	0	0	TA10110
Grp311	1	1	0	0	TA10112
Grp312	1	1	0	0	TA10113
Grp313	1	1	0	0	TA10114
Grp314	1	1	0	0	TA10115
Grp315	1	1	0	0	TA10116
Grp316	1	1	0	0	TA10118
Grp317	3	3	0	0	TA10119, TA10187, TA10188
Grp318	1	1	0	0	TA10120
Grp319	1	1	0	0	TA10121
Grp320	1	1	0	0	TA10122
Grp321	1	1	0	0	TA10125
Grp322	2	2	0	0	TA10126, TA10127
Grp323	1	1	0	0	TA10128
Grp324	1	1	0	0	TA10132
Grp325	1	1	0	0	TA10134
Grp326	1	1	0	0	TA10135
Grp327	1	1	0	0	TA10136
Grp328	1	1	0	0	TA10137
Grp329	1	1	0	0	TA10138
Grp330	1	1	0	0	TA10142
Grp331	3	3	0	0	TA10143, TA10144, TA10145
Grp332	1	1	0	0	TA10155
Grp333	1	1	0	0	TA10156

Grp334	1	1	0	0	TA10157
Grp335	1	1	0	0	TA10158
Grp336	1	1	0	0	TA10159
Grp337	1	1	0	0	TA10160
Grp338	1	1	0	0	TA10161
Grp339	1	1	0	0	TA10162
Grp340	1	1	0	0	TA10163
Grp341	1	1	0	0	TA10164
Grp342	2	2	0	0	TA10165, TA10185
Grp343	2	2	0	0	TA10166, TA10167
Grp344	1	1	0	0	TA10168
Grp345	1	1	0	0	TA10169
Grp346	1	1	0	0	TA10170
Grp347	1	1	0	0	TA10171
Grp348	2	2	0	0	TA10172, TA10173
Grp349	1	1	0	0	TA10174
Grp350	1	1	0	0	TA10175
Grp351	1	1	0	0	TA10176
Grp352	1	1	0	0	TA10177
Grp353	1	1	0	0	TA10178
Grp354	1	1	0	0	TA10179
Grp355	1	1	0	0	TA10180
Grp356	1	1	0	0	TA10182
Grp357	1	1	0	0	TA10183
Grp358	1	1	0	0	TA10184
Grp359	1	1	0	0	TA10186
Grp360	1	1	0	0	TA10189
Grp361	1	1	0	0	TA10190
Grp362	1	1	0	0	TA10191
Grp363	1	1	0	0	TA10192
Grp364	1	1	0	0	TA10193
Grp365	1	1	0	0	TA10194
Grp366	2	2	0	0	TA10195, TA10198
Grp367	1	1	0	0	TA10196
Grp368	1	1	0	0	TA10197
Grp369	1	1	0	0	TA10199
Grp370	1	1	0	0	TA10200
Grp371	1	1	0	0	TA10202
Grp372	1	1	0	0	TA10210
Grp373	1	1	0	0	TA10211
Grp374	1	1	0	0	TA10212
Grp375	1	1	0	0	TA10213
Grp376	3	3	0	0	TA10291, TA10292, TA10314
Grp377	3	3	0	0	TA10293, TA10296, TA10304
Grp378	7	7	0	0	TA10294, TA10295, TA10297, TA10298, TA10299, TA10300, TA10303
Grp379	2	2	0	0	TA10301, TA10302
Grp380	3	3	0	0	TA10305, TA10309, TA10311
Grp381	2	2	0	0	TA10306, TA10308
Grp382	1	1	0	0	TA10307
Grp383	1	1	0	0	TA10310

Grp384	2	2	0	0	TA10312, TA10313
Grp385	2	2	0	0	TA10315, TA10319
Grp386	2	2	0	0	TA10316, TA10317
Grp387	1	1	0	0	TA10318
Grp388	1	1	0	0	TA10320
Grp389	1	1	0	0	TA10321
Grp390	1	1	0	0	TA10322
Grp391	1	1	0	0	TA10323
Grp392	2	2	0	0	TA10324, TA10325
Grp393	1	1	0	0	TA10326
Grp394	1	1	0	0	TA10327
Grp395	1	1	0	0	TA10328
Grp396	1	1	0	0	TA10329
Grp397	2	2	0	0	TA10330, TA10331
Grp398	1	1	0	0	TA10417
Grp399	1	1	0	0	TA10836
Grp400	1	1	0	0	TA10837
Grp401	1	1	0	0	TA10838
Grp402	1	1	0	0	TA10839
Grp403	2	2	0	0	TA10869, TA1671
Grp404	1	1	0	0	TA10871
Grp405	1	1	0	0	TA10872
Grp406	1	1	0	0	TA10918
Grp407	2	2	0	0	TA10919, TA10921
Grp408	1	1	0	0	TA10920
Grp409	1	1	0	0	TA10922
Grp410	1	1	0	0	TA10923
Grp411	1	1	0	0	TA10924
Grp412	1	1	0	0	TA10925
Grp413	1	1	0	0	TA10926
Grp414	1	1	0	0	TA10927
Grp415	2	2	0	0	TA10928, TA10929
Grp416	1	1	0	0	TA10930
Grp417	1	1	0	0	TA10931
Grp418	1	1	0	0	TA10932
Grp419	2	2	0	0	TA10933, TA10934
Grp420	1	1	0	0	TA10935
Grp421	1	1	0	0	TA10936
Grp422	4	4	0	0	TA10937, TA1653, TA1667, TA1678
Grp423	1	1	0	0	TA10938
Grp424	1	1	0	0	TA10939
Grp425	1	1	0	0	TA10940
Grp426	1	1	0	0	TA10941
Grp427	1	1	0	0	TA10942
Grp428	1	1	0	0	TA10943
Grp429	1	1	0	0	TA10944
Grp430	1	1	0	0	TA10945
Grp431	1	1	0	0	TA10946
Grp432	1	1	0	0	TA10947
Grp433	2	2	0	0	TA10948, TA10949
Grp434	1	1	0	0	TA10950

Grp435	2	2	0	0	TA10951, TA10952
Grp436	1	1	0	0	TA10953
Grp437	2	2	0	0	TA10954, TA10958
Grp438	1	1	0	0	TA10955
Grp439	1	1	0	0	TA10956
Grp440	1	1	0	0	TA10957
Grp441	2	2	0	0	TA10959, TA10960
Grp442	1	1	0	0	TA10961
Grp443	1	1	0	0	TA11020
Grp444	1	1	0	0	TA11021
Grp445	1	1	0	0	TA1586
Grp446	1	1	0	0	TA1592
Grp447	1	1	0	0	TA1606
Grp448	1	1	0	0	TA1612
Grp449	1	1	0	0	TA1616
Grp450	1	1	0	0	TA1619
Grp451	1	1	0	0	TA1630
Grp452	1	1	0	0	TA1633
Grp453	1	1	0	0	TA1640
Grp454	1	1	0	0	TA1641
Grp455	4	4	0	0	TA1643, TA2524, TA2525, TA2568
Grp456	1	1	0	0	TA1652
Grp457	1	1	0	0	TA1656
Grp458	1	1	0	0	TA1662
Grp459	1	1	0	0	TA1666
Grp460	1	1	0	0	TA1668
Grp461	1	1	0	0	TA1669
Grp462	1	1	0	0	TA1670
Grp463	1	1	0	0	TA1672
Grp464	1	1	0	0	TA1673
Grp465	1	1	0	0	TA1674
Grp466	1	1	0	0	TA1677
Grp467	1	1	0	0	TA1679
Grp468	1	1	0	0	TA1682
Grp469	1	1	0	0	TA1683
Grp470	1	1	0	0	TA1684
Grp471	3	3	0	0	TA1685, TA1713, TA1714
Grp472	1	1	0	0	TA1691
Grp473	1	1	0	0	TA1698
Grp474	1	1	0	0	TA1712
Grp475	1	1	0	0	TA1715
Grp476	4	4	0	0	TA1716, TA2500, TA2501, TA2560
Grp477	1	1	0	0	TA1718
Grp478	1	1	0	0	TA2119
Grp479	1	1	0	0	TA2373
Grp480	1	1	0	0	TA2376
Grp481	1	1	0	0	TA2383
Grp482	1	1	0	0	TA2384
Grp483	1	1	0	0	TA2389
Grp484	1	1	0	0	TA2393
Grp485	1	1	0	0	TA2395

Grp486	1	1	0	0	TA2396
Grp487	1	1	0	0	TA2398
Grp488	1	1	0	0	TA2399
Grp489	1	1	0	0	TA2400
Grp490	1	1	0	0	TA2401
Grp491	1	1	0	0	TA2403
Grp492	1	1	0	0	TA2404
Grp493	1	1	0	0	TA2405
Grp494	1	1	0	0	TA2409
Grp495	1	1	0	0	TA2410
Grp496	1	1	0	0	TA2414
Grp497	1	1	0	0	TA2417
Grp498	1	1	0	0	TA2418
Grp499	1	1	0	0	TA2419
Grp500	1	1	0	0	TA2426
Grp501	1	1	0	0	TA2429
Grp502	1	1	0	0	TA2430
Grp503	1	1	0	0	TA2432
Grp504	1	1	0	0	TA2433
Grp505	1	1	0	0	TA2438
Grp506	1	1	0	0	TA2440
Grp507	1	1	0	0	TA2441
Grp508	1	1	0	0	TA2443
Grp509	1	1	0	0	TA2444
Grp510	1	1	0	0	TA2459
Grp511	1	1	0	0	TA2460
Grp512	1	1	0	0	TA2461
Grp513	1	1	0	0	TA2466
Grp514	1	1	0	0	TA2471
Grp515	1	1	0	0	TA2479
Grp516	1	1	0	0	TA2480
Grp517	1	1	0	0	TA2483
Grp518	2	2	0	0	TA2484, TA2485
Grp519	1	1	0	0	TA2489
Grp520	2	2	0	0	TA2492, TA2493
Grp521	1	1	0	0	TA2497
Grp522	1	1	0	0	TA2498
Grp523	2	2	0	0	TA2502, TA2503
Grp524	1	1	0	0	TA2504
Grp525	1	1	0	0	TA2511
Grp526	1	1	0	0	TA2512
Grp527	1	1	0	0	TA2513
Grp528	2	2	0	0	TA2514, TA2515
Grp529	2	2	0	0	TA2516, TA2517
Grp530	1	1	0	0	TA2518
Grp531	1	1	0	0	TA2519
Grp532	1	1	0	0	TA2526
Grp533	1	1	0	0	TA2529
Grp534	1	1	0	0	TA2531
Grp535	1	1	0	0	TA2533
Grp536	1	1	0	0	TA2542

Grp537	2	2	0	0	TA2544, TA2555
Grp538	1	1	0	0	TA2545
Grp539	1	1	0	0	TA2546
Grp540	2	2	0	0	TA2547, TA2548
Grp541	2	2	0	0	TA2550, TA2551
Grp542	1	1	0	0	TA2552
Grp543	1	1	0	0	TA2553
Grp544	2	2	0	0	TA2556, TA2557
Grp545	1	1	0	0	TA2558
Grp546	1	1	0	0	TA2559
Grp547	1	1	0	0	TA2561
Grp548	2	2	0	0	TA2562, TA2565
Grp549	1	1	0	0	TA2564
Grp550	1	1	0	0	TA2566
Grp551	1	1	0	0	TA2571
Grp552	2	2	0	0	TA2572, TA2573
Grp553	1	1	0	0	TA2574
Grp554	1	1	0	0	TA2576
Grp555	1	1	0	0	TA2577
Grp556	1	1	0	0	TA2578
Grp557	1	1	0	0	TA2579
Grp558	1	1	0	0	TA2581
Grp559	1	1	0	0	TA2582
Grp560	1	1	0	0	TA2583
Grp561	1	1	0	0	TA2585
Grp562	1	1	0	0	TA2586
Grp563	1	1	0	0	TA2587
Grp564	1	1	0	0	TA3429

Table B.3. Imputed posterior probabilities for each accession with missing geographical information.

Highlighted in red are the probabilities that place an accession in a specific group with high probability. Column one is accession names, and column two is inferred origin of country based on the posterior probability. Starting column three are the three letter country codes- AFG=Afghanistan, ARM=Armenia, AZB=Azerbaijan, CHI=China, GEO=Georgia, IRN=Iran, KYR=Kyrgyzstan, PAK=Pakistan, RUS=Russia, SYR=Syria, TAJ=Tajikistan, TKY=Turkey, TKM=Turkmenistan, and UZB=Uzbekistan. UNK=Unknown in the second column.

		Imputed posterior probabilities for each geographical region													
Accessions	Inferred	AFG	ARM	AZB	CHI	GEO	IRN	KYR	PAK	RUS	SYR	TAJ	TKY	TKM	UZB
TA10148	TKY	0.028	0.049	0	0.057	0.06	0	0	0.004	0	0	0	0.801	0	0
TA10417	IRN	0	0	0.086	0	0	0.914	0	0	0	0	0	0	0	0
TA11021	AZB	0	0	0.996	0	0	0.004	0	0	0	0	0	0	0	0
TA1577	UNK	0	0.008	0	0	0	0	0.014	0	0.001	0.147	0.043	0.301	0.483	0.003
TA1578	AFG	1	0	0	0	0	0	0	0	0	0	0	0	0	0
TA1579	AFG	1	0	0	0	0	0	0	0	0	0	0	0	0	0
TA1580	AFG	1	0	0	0	0	0	0	0	0	0	0	0	0	0
TA1595	AFG	0.756	0	0	0	0	0	0	0	0	0	0.199	0	0.045	0
TA1596	AFG	0.716	0	0	0	0	0	0	0	0	0	0.11	0	0.174	0
TA1597	AFG	0.771	0	0	0	0	0	0	0	0	0.001	0.18	0.003	0.046	0
TA1605	IRN	0	0	0.046	0	0	0.954	0	0	0	0	0	0	0	0
TA1650	AFG	0.988	0	0	0	0	0	0	0	0	0	0.012	0	0	0
TA1688	AFG	0.886	0	0	0	0	0	0	0	0	0	0	0	0.114	0
TA1689	AFG	0.929	0	0	0	0	0	0.001	0	0	0.017	0	0	0.053	0
TA1691	IRN	0	0	0	0	0	1	0	0	0	0	0	0	0	0
TA1695	IRN	0	0	0	0	0	0.999	0	0	0	0	0	0	0	0
TA1696	IRN	0	0	0.011	0	0	0.989	0	0	0	0	0	0	0	0
TA1697	AFG	0.951	0.008	0	0	0	0	0	0	0	0.04	0.002	0	0	0
TA1707	UNK	0	0	0	0.018	0	0	0.045	0.022	0	0	0	0.25	0.455	0.209
TA1708	AFG	0.993	0	0	0	0	0	0	0	0	0	0.006	0	0	0
TA1709	AFG	1	0	0	0	0	0	0	0	0	0	0	0	0	0
TA1712	AFG	0.684	0	0	0	0	0	0	0	0	0	0.199	0	0.117	0
TA2118	AFG	1	0	0	0	0	0	0	0	0	0	0	0	0	0
TA2370	TKM	0.158	0	0	0	0	0	0	0	0	0	0	0	0.842	0
TA2373	TKY	0	0.001	0	0.031	0.002	0	0	0.132	0.013	0	0.003	0.656	0.162	0
TA2378_L1	IRN	0	0	0	0	0	1	0	0	0	0	0	0	0	0

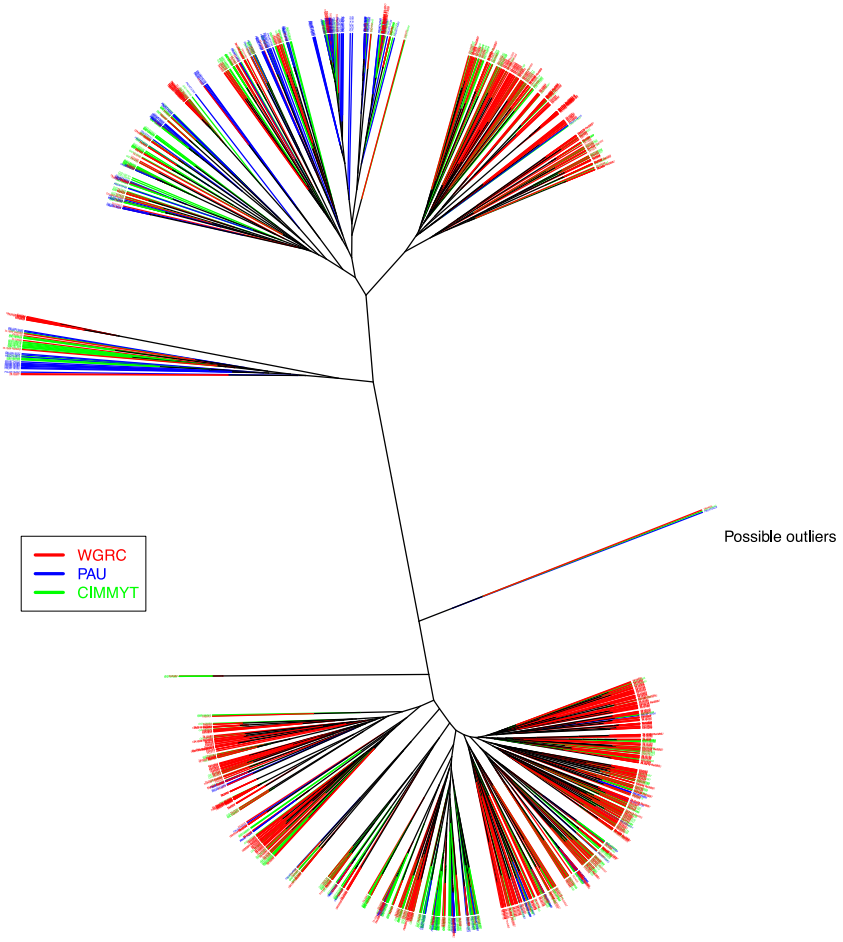

Figure B.1. Cluster analysis showing *Ae. tauschii* grouping and possible outliers.

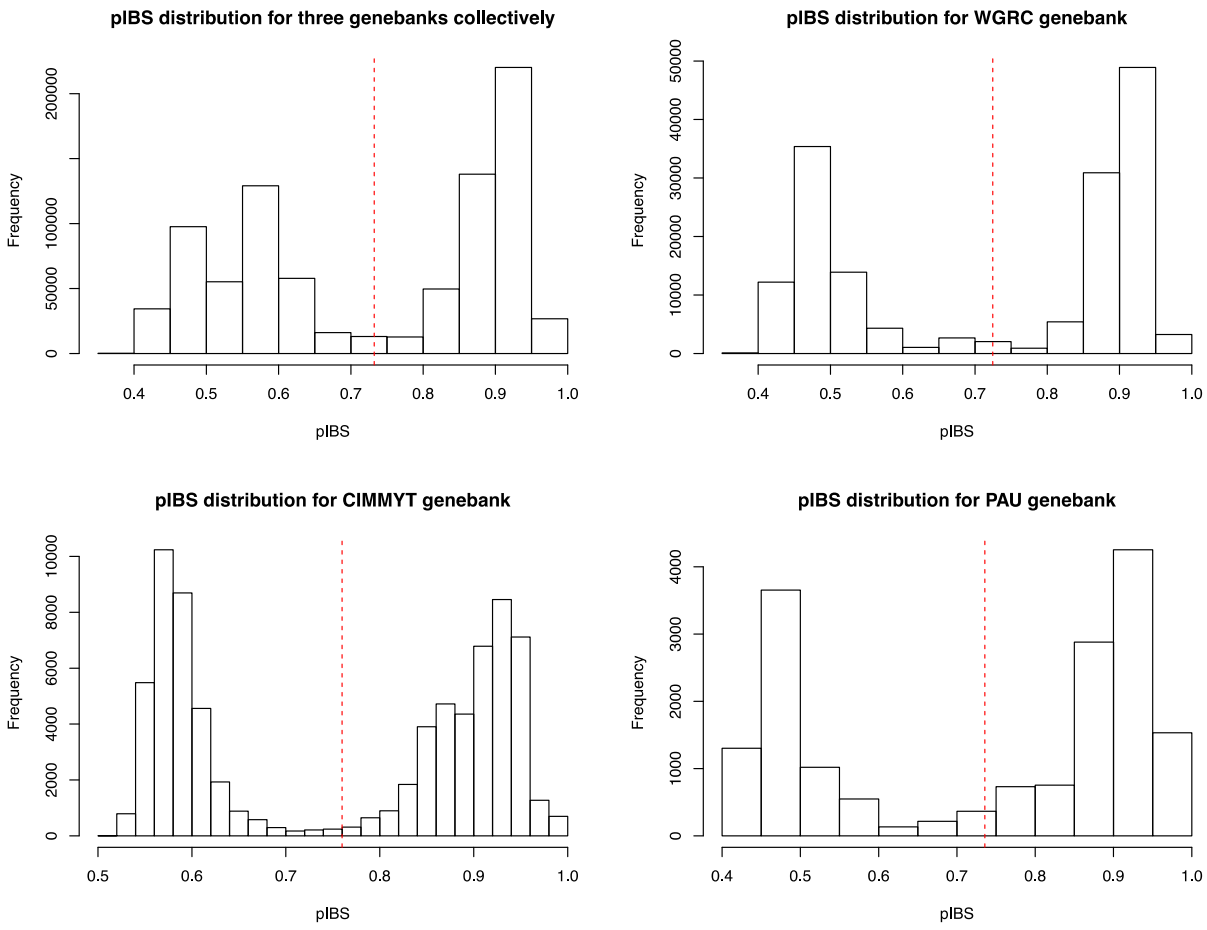

Figure B.2. Percent identity by state (pIBS) coefficient distributions for all three genebanks collectively and individually.

Appendix C - Supplementary material Chapter 3

This appendix contains supplementary figures and tables for Chapter 3.

Table C.1. List of PowerCore accessions.

MC in second column represents if the accession is the part of MiniCore; SR = stem rust race TTKSK, LR = leaf rust composite, HF = Hessian fly biotype L; R = resistant, S = susceptible, MR = moderately resistant, H = heterozygous.

TA	CORE	ORIGIN	LATITUDE	LONGITUDE	ALTITUDE	SR	LR	HF
TA10099	MC	Armenia	40.139722	44.527778	995 m	R	-	-
TA10106	MC	Kyrgyzstan	40.7081	72.878838	743 m	R	-	-
TA10108	MC	Tajikistan	39.215745	70.526493	-	S	-	-
TA10124	MC	Uzbekistan	41.266667	69.216667	433 m	R	-	-
TA10141	MC	PR Chi-	43.521599	82.692992	-	S	-	-
TA10144	MC	Syria	35.625	38.75559998	370 m	S	-	-
TA10162	MC	Turkmenistan	38	58.1167	610 m	S	-	-
TA10179	MC	Turkmenistan	38.25	56.33333333	895 m	R	-	-
TA10210	MC	Uzbekistan	41.23333333	71.65183333	930 m	S	-	-
TA10212	MC	Uzbekistan	40.31666667	71.76716667	773 m	S	-	-
TA10330	MC	Kyrgyzstan	42.05745	79.07959	2550 m	S	-	-
TA1578	MC	Unknown	-	-	-	-	R	H
TA1596	MC	Unknown	-	-	-	S	MR	H
TA1605	MC	Unknown	-	-	-	R	R	-
TA1631	MC	Afghanistan	35.71666667	64.9	915 m	-	MR	S
TA1651	MC	Iran	36.846659	54.440002	130 m	R	R	R
TA1665	MC	Azerbaijan	40.08333333	49.4	650 m	R	R	R
TA1666	MC	Azerbaijan	40.08333333	49.4	630 m	R	R	R
TA1667	MC	Azerbaijan	40.08333333	49.4	630 m	S	R	R
TA1669	MC	Azerbaijan	40.08333333	49.4	780 m	-	R	R
TA1694	MC	Turkmenistan	38.483333	56.3	450 m	S	R	S
TA1707	MC	Sweden	-	-	-	H	R	R
TA2374	MC	Pakistan	30.132063	66.96167	1690 m	S	S	S
TA2376	MC	Iran	35.681841	52.514648	2010 m	R	S	S
TA2378	MC	Iran	36.957574	50.594788	-15 m	R	R	S
TA2395	MC	Afghanistan	34.019657	68.729095	2075 m	-	S	S
TA2413	MC	Afghanistan	36.176128	68.377533	1040 m	S	S	S
TA2431	MC	Afghanistan	35.906849	64.697113	860 m	S	S	R
TA2435	MC	Afghanistan	35.784399	64.371643	1100 m	-	S	R
TA2448	MC	Iran	35.89795	50.977592	1600 m	-	S	R
TA2458	MC	Iran	37.399074	55.500183	100 m	-	R	R
TA2468	MC	Iran	36.590171	52.090645	-20 m	H	R	R
TA2474	MC	Iran	37.151561	50.245972	-24 m	R	S	R
TA2485	MC	Iran	38.101063	48.126984	2240 m	R	S	S
TA2488	MC	Iran	37.084762	45.479279	1270 m	S	S	R
TA2508	MC	Turkey	38.957273	43.659668	1780 m	S	S	R
TA2514	MC	Iran	38.90279	45.032959	1100 m	S	S	R
TA2536	MC	Afghanistan	36.469889	69.866867	1300 m	S	S	H

TA2545	MC	Afghanistan	36.955105	71.012878	1460 m	S	S	R
TA2586	MC	Georgia	42.017162	44.141693	630 m	R	S	H
TA2493	-	Iran	38.458966	45.608368	1170 m	S	S	R
TA2453	-	Iran	36.966901	54.956703	78 m	-	S	R
TA2521	-	Iran	37.027773	48.004074	1235 m	S	S	S
TA1625	-	Azerbaijan	40.570154	48.747711	-	-	S	H
TA10181	-	Turkmenistan	40.03333333	52.96666667	26 m	S	-	-
TA2539	-	Afghanistan	37.069971	71.141968	2060 m	S	S	S
TA10945	-	Azerbaijan	40.58615	48.4158	370 m	-	-	-
TA10960	-	Azerbaijan	39.543833	45.260167	2018 m	-	-	-
TA1626	-	Turkmenistan	38.416667	57.4	1200 m	-	S	H
TA1670	-	Azerbaijan	40.983333	47.833333	600 m	H	R	R
TA1699	-	Russia	42.133333	47.1	1850 m	S	S	H
TA10192	-	Uzbekistan	39.96666667	37.5	733 m	S	-	-
TA2480	-	Iran	37.903574	48.90152	15 m	R	S	H
TA10932	-	Azerbaijan	40.50005	48.980183	631 m	-	-	-
TA10317	-	Tajikistan	38.583333	68.8	800 m	H	-	-
TA2512	-	Iran	39.248207	44.88327	900 m	S	S	H
TA10131	-	Armenia	39.507882	46.338615	1800 m	S	-	-
TA2565	-	Azerbaijan	40.570154	48.747711	-	-	S	S
TA10136	-	PR Chi-	43.814711	82.468872	730 m	S	-	-
TA2530	-	Iran	36.90598	50.657959	0 m	S	R	H
TA2462	-	Iran	36.605055	53.22464	43 m	-	R	-
TA10142	-	Syria	36.71889877	40.14749908	340 m	R	-	-
TA10310	-	Tajikistan	38.566667	69.316667	1600 m	H	-	-
TA2549	-	Afghanistan	37.030514	71.012192	1750 m	S	S	H
TA2475	-	Iran	37.188767	50.151215	-20 m	-	S	R
TA2559	-	Afghanistan	36.390335	68.89801	480 m	S	S	H
TA1612	-	Former USSR	-	-	-	S	-	-
TA2466	-	Iran	36.695402	52.621078	-22 m	-	S	R
TA2540	-	Afghanistan	37.069971	71.141968	2060 m	S	S	S
TA10167	-	Turkmenistan	38.341656	56.297379	710 m	S	-	-
TA2575	-	Armenia	40.33503	44.264774	1450 m	S	S	S
TA2497	-	Iran	36.65	51.42	30 m	-	S	H
TA2455	-	Iran	36.918607	54.884949	145 m	S	S	R
TA2479	-	Iran	37.533688	49.270935	-25 m	S	S	S
TA2510	-	Turkey	38.985033	43.558044	1740 m	-	S	R
TA10169	-	Turkmenistan	38.2	56.2	470 m	S	-	-
TA2427	-	Afghanistan	35.918389	64.767151	850 m	S	S	H
TA1655	-	Afghanistan	36.390335	68.89801	518 m	-	S	S
TA10198	-	Uzbekistan	40.88333333	71.1	559 m	S	-	-
TA2464	-	Iran	36.694851	53.536377	10 m	-	S	R
TA10127	-	Georgia	41.838611	44.733889	470 m	S	-	-
TA2419	-	Afghanistan	36.188875	68.306808	1330 m	S	S	R
TA10182	-	Turkmenistan	37.85	58.36333333	280 m	S	-	-
TA10145	-	Syria	35.62559891	38.75859833	-	S	-	-
TA10193	-	Uzbekistan	39.5	67.38333333	1228 m	S	-	-
TA2532	-	Afghanistan	34.550822	69.034481	2020 m	-	S	-
TA10918	-	Georgia	41.60986	44.80171	521 m	-	-	-
TA10303	-	Tajikistan	38.525556	68.547222	450 m	S	-	-
TA2465	-	Iran	36.695402	52.621078	-22 m	-	S	R

TA2469	-	Iran	36.590171	52.090645	-20 m	-	S	R
TA1657	-	Azerbaijan	40.75	48.73333333	410 m	S	MR	S
TA10104	-	Georgia	41.645472	44.854343	580 m	S	-	-
TA10296	-	Tajikistan	38.525556	68.547222	450 m	-	-	-
TA10921	-	Georgia	41.60986	44.80171	521 m	-	-	-
TA1679	-	Azerbaijan	40.63333333	48.61666667	780 m	-	S	S
TA10185	-	Turkmenistan	38.5	56.83333333	1463 m	S	-	-
TA2407	-	Afghanistan	36.137875	68.513489	675 m	S	S	S
TA1641	-	Iran	36.611118	53.231163	33 m	S	R	S
TA10188	-	Turkmenistan	38.48805556	56.70416667	1073 m	R	-	-
TA2382	-	Pakistan	30.206861	66.967163	1645 m	H	S	H
TA2377	-	Iran	36.692373	53.475609	8 m	H	S	R
TA1668	-	Azerbaijan	40.08333333	49.4	800 m	R	S	R
TA2449	-	Iran	36.692373	53.475609	8 m	-	R	R
TA2372	-	Afghanistan	35.724218	63.484497	-	-	S	H
TA2515	-	Iran	38.90279	45.032959	1100 m	S	S	H
TA2426	-	Afghanistan	35.918389	64.767151	850 m	S	S	H
TA2461	-	Iran	35.724218	52.663651	2025 m	S	S	S

Figure C.1. Geographical distribution of previously collected *Ae. tauschii* accessions from '50s and '60s (red dots) and newly collected 2012 accessions (blue dots).

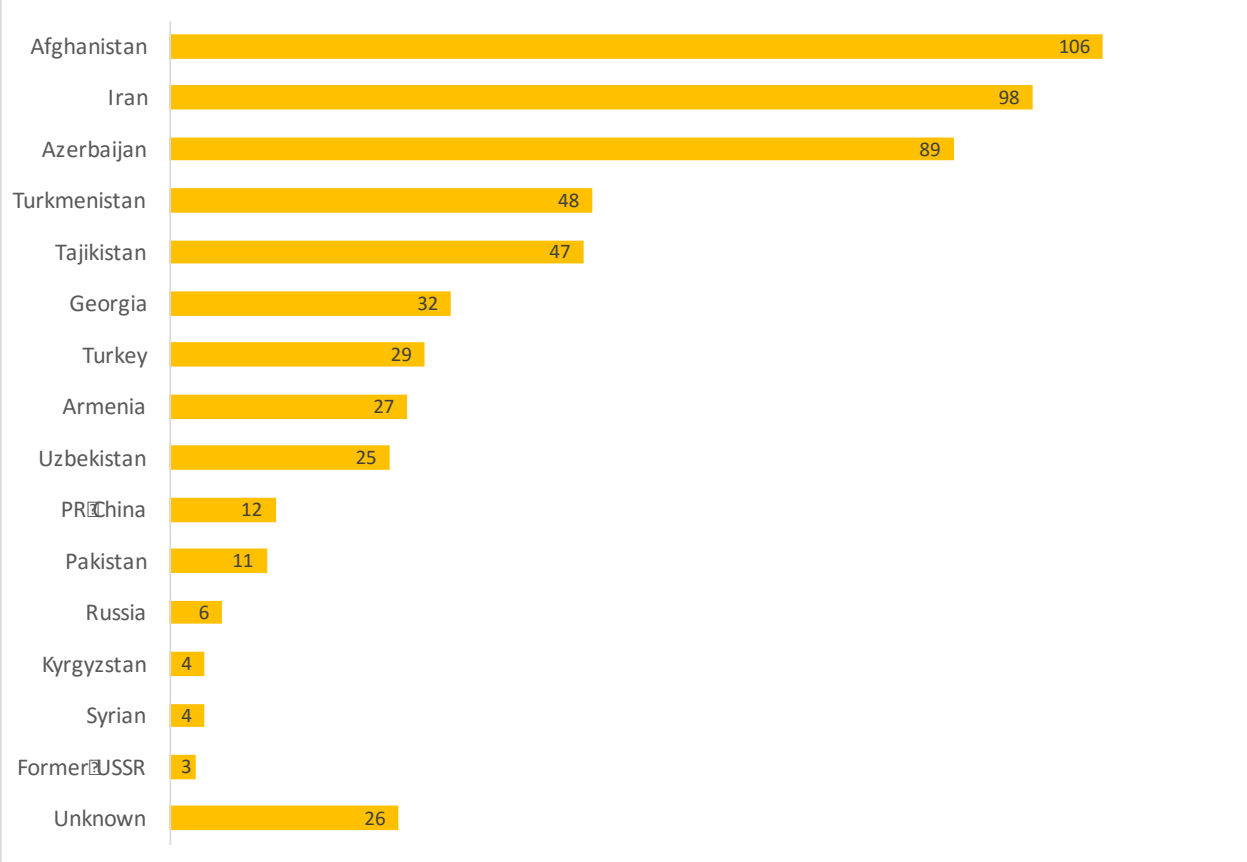

Figure C.2. Country wise distribution of *Ae. tauschii* accessions.

Countries on the y-axis are ordered according to the number of accessions contributed (higher to lower).

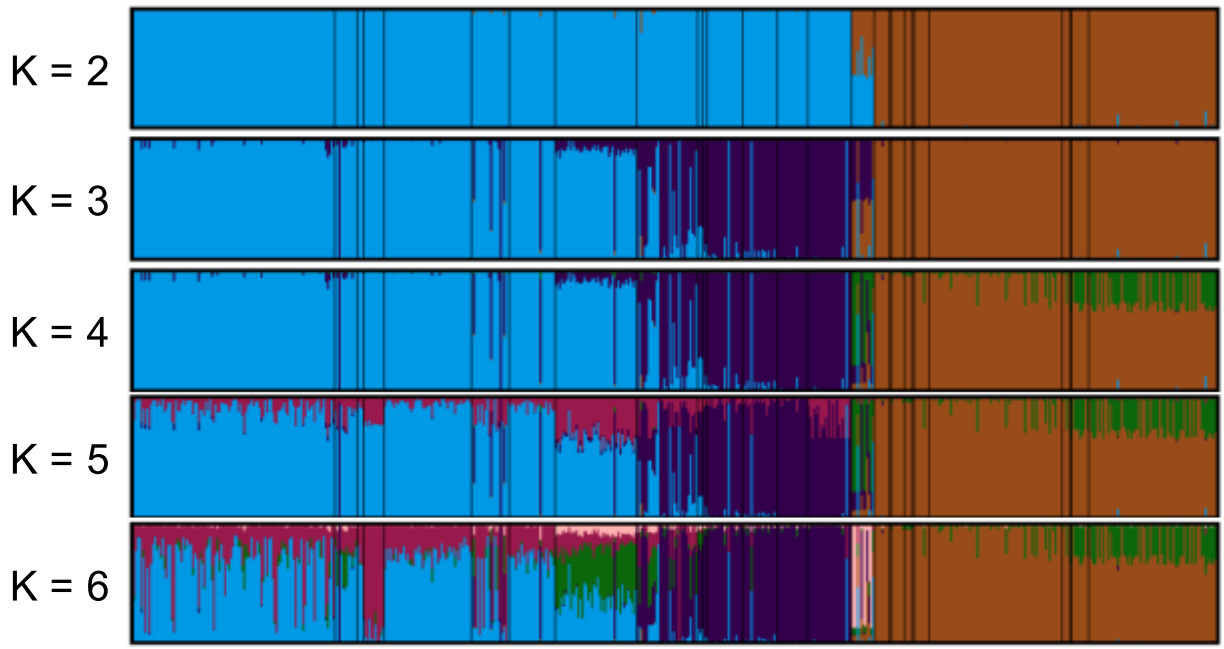

Figure C.3. STRUCTURE analysis for all *Ae. tauschii* accessions for K=2 to K=6.

Each vertical bar represents an individual. A bar with only a single color represents its ancestry to a single population, and a mixture of colors represents admixture from different populations.

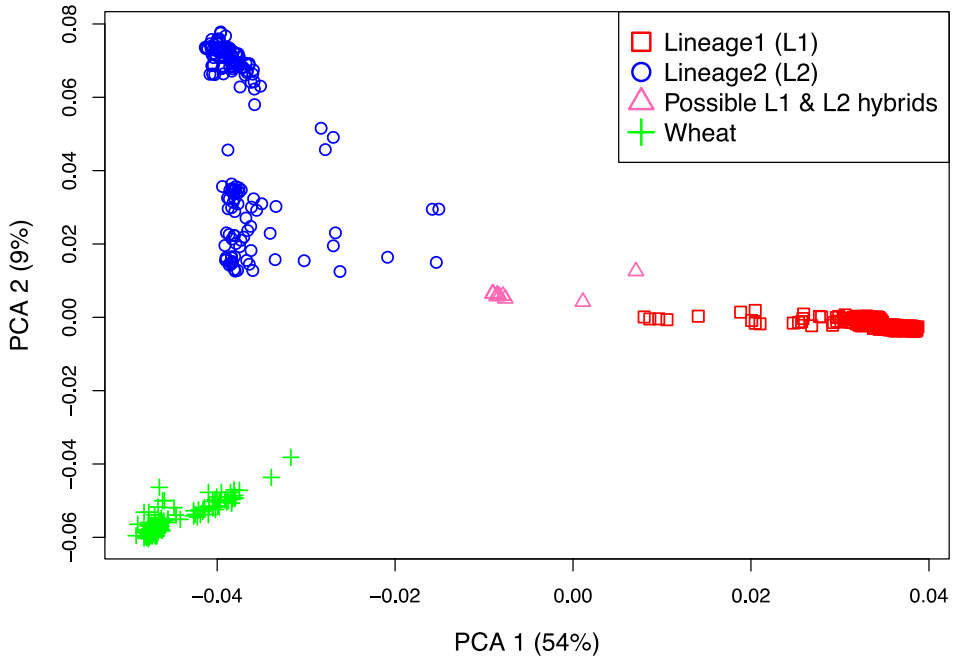

Figure C.4. PCA showing the clustering of *Ae. tauschii* L1, L2 and wheat.

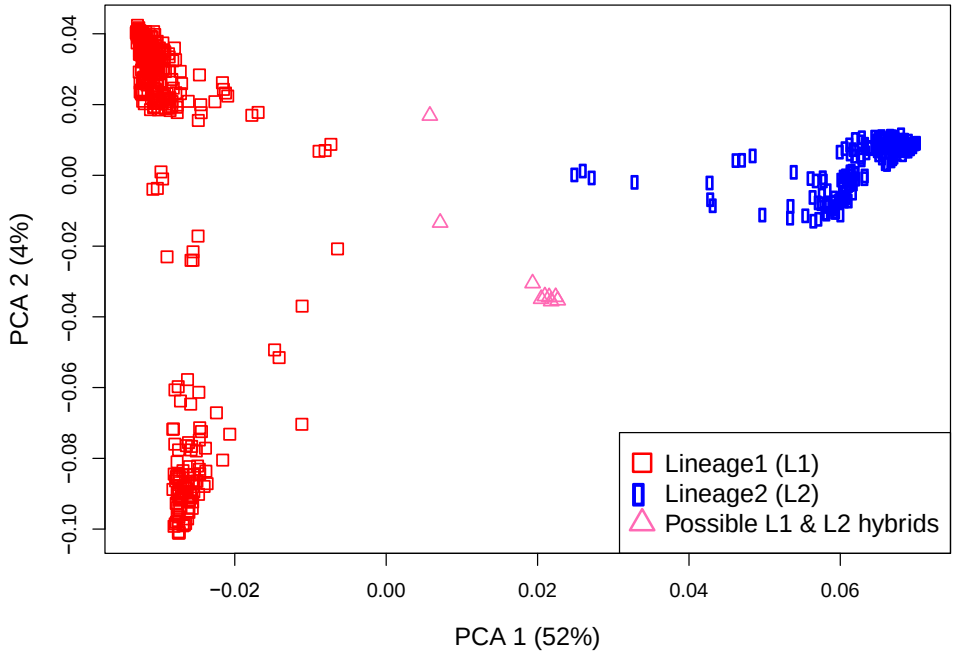

Figure C.5. PCA showing the differentiation of *Ae. tauschii* L1 and L2. Putative hybrids are shown in the middle.

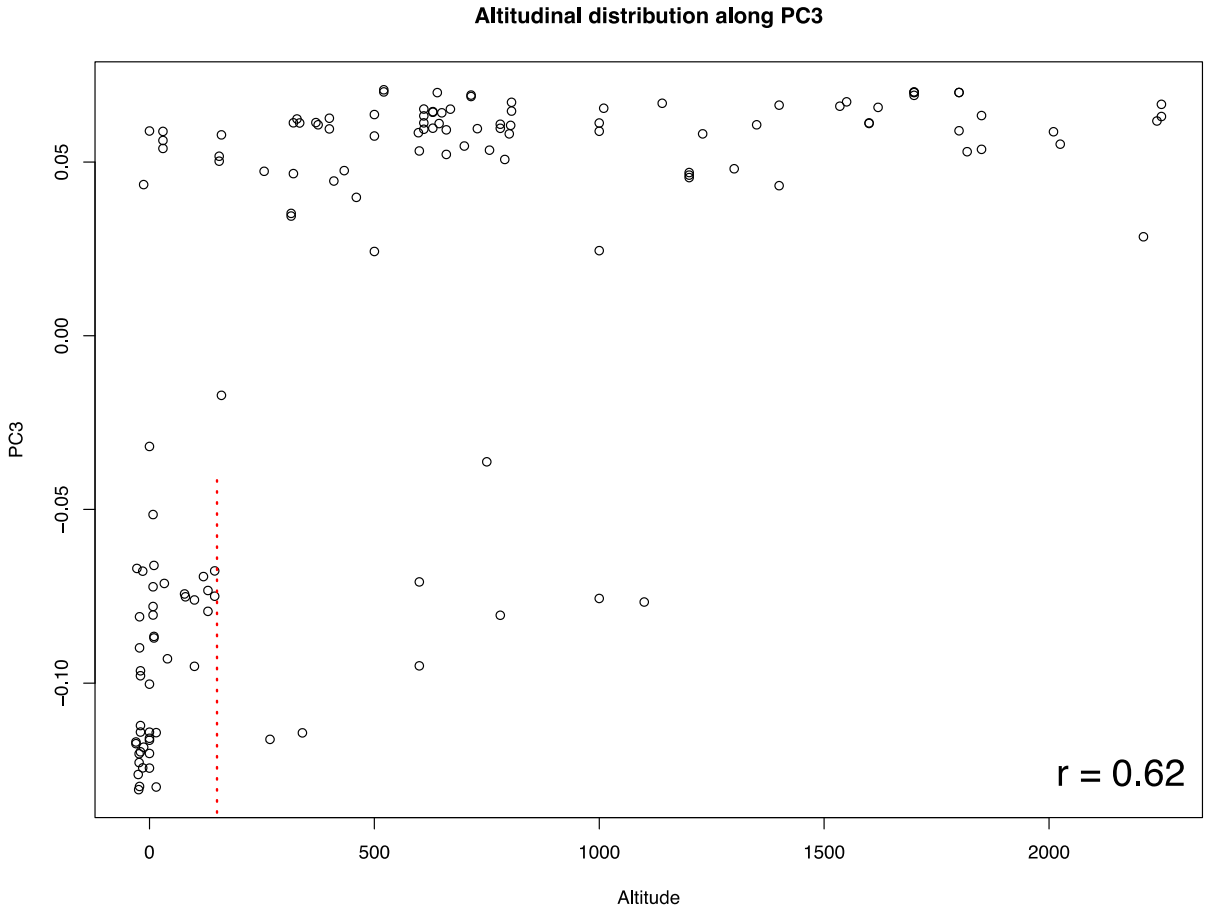

Figure C.6. Scatterplot showing the relationship between L2 altitude and PC3.

Strong correlation between two variables is evident suggesting that PC3 is separating out lower and higher altitude accessions. Correlation coefficient is shown at the bottom right corner. Vertical red dotted line marks the 150m altitude separating these clusters.

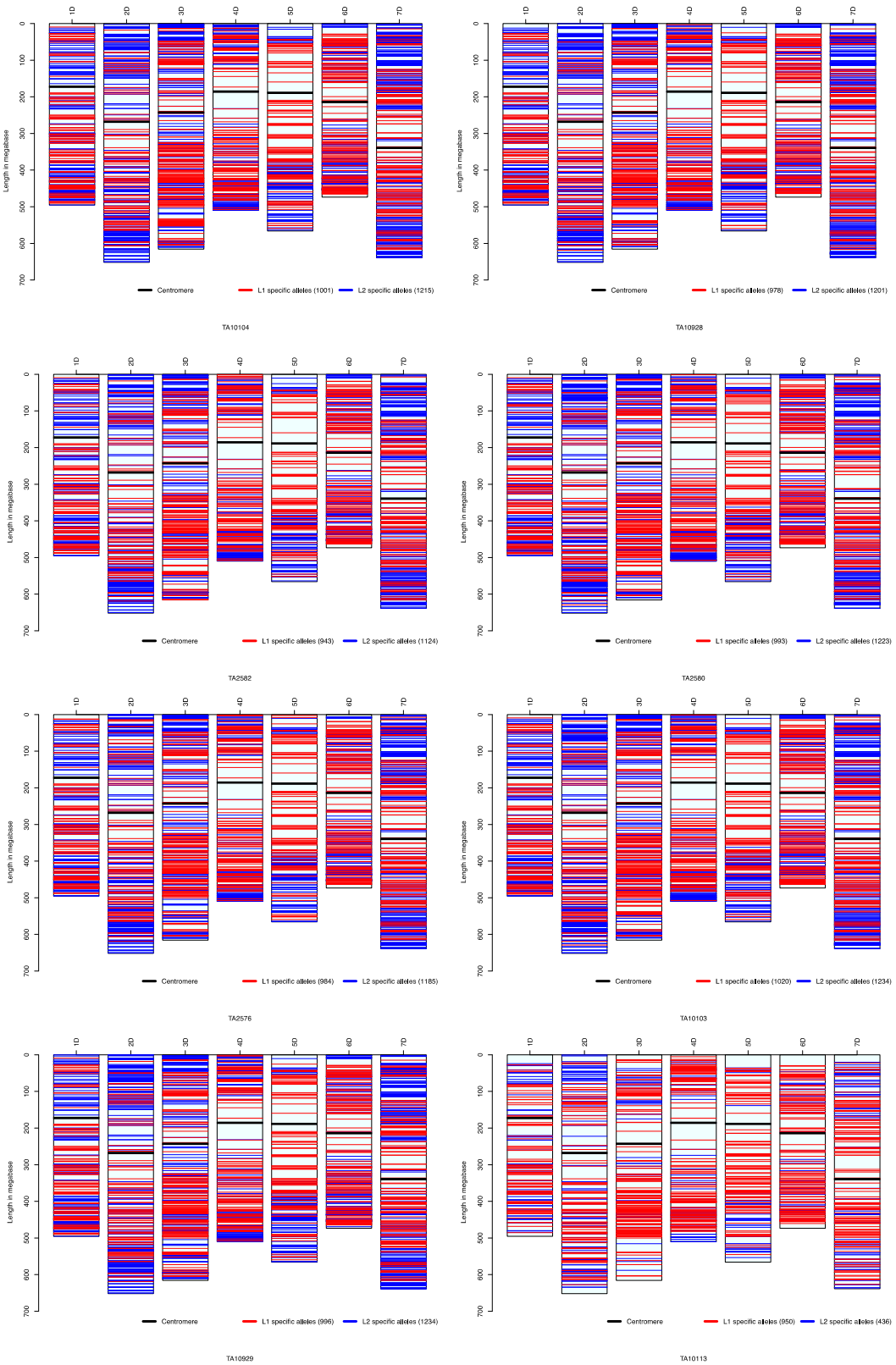

Figure C.7. Distribution of lineage specific alleles for putative hybrid samples. Red color represents L1 specific alleles, and blue represents L2 specific alleles.

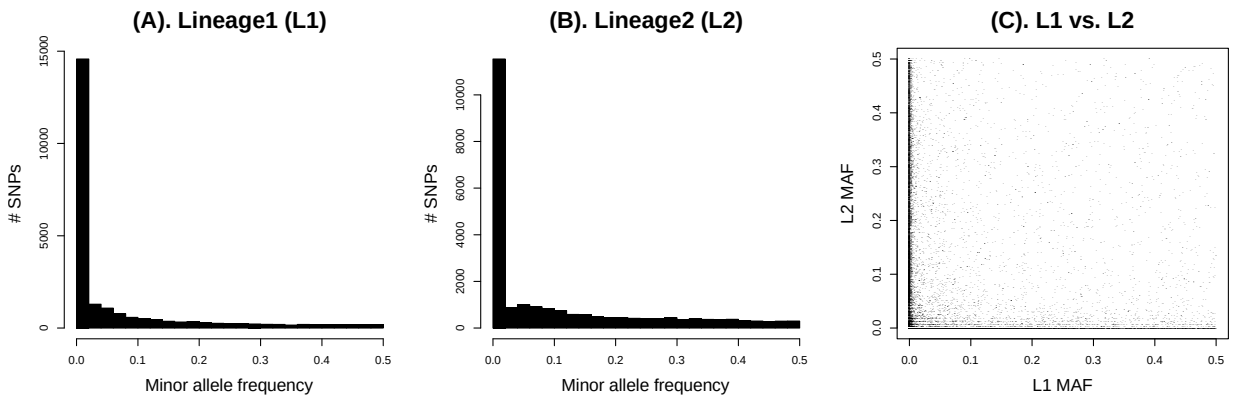

Figure C.8. Minor allele frequency plots for L1, L2, and joint L1 and L2 MAF distribution.

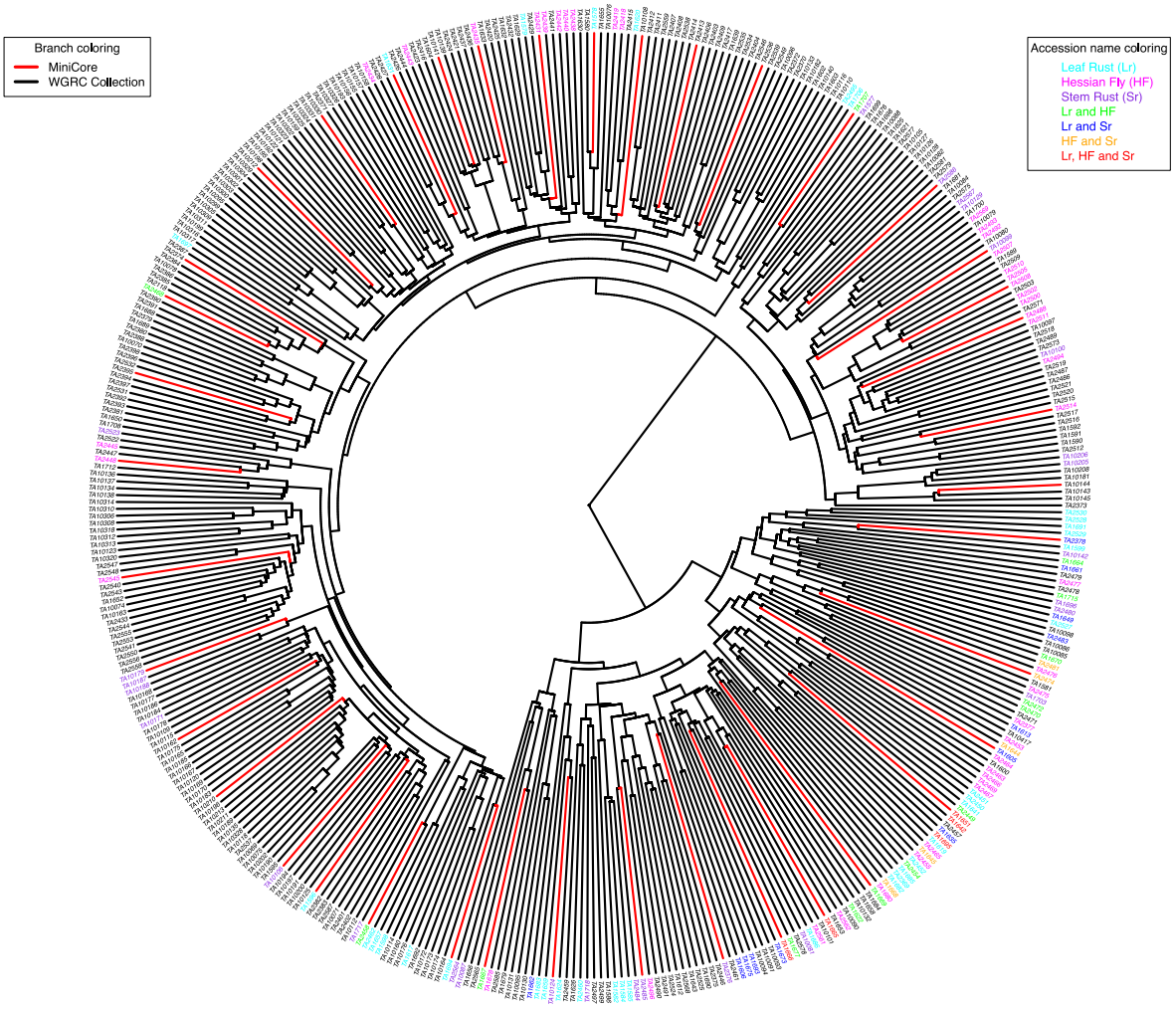

Figure C.9. Distribution of MiniCore (red branches) within the whole *Ae. tauschii* collection.

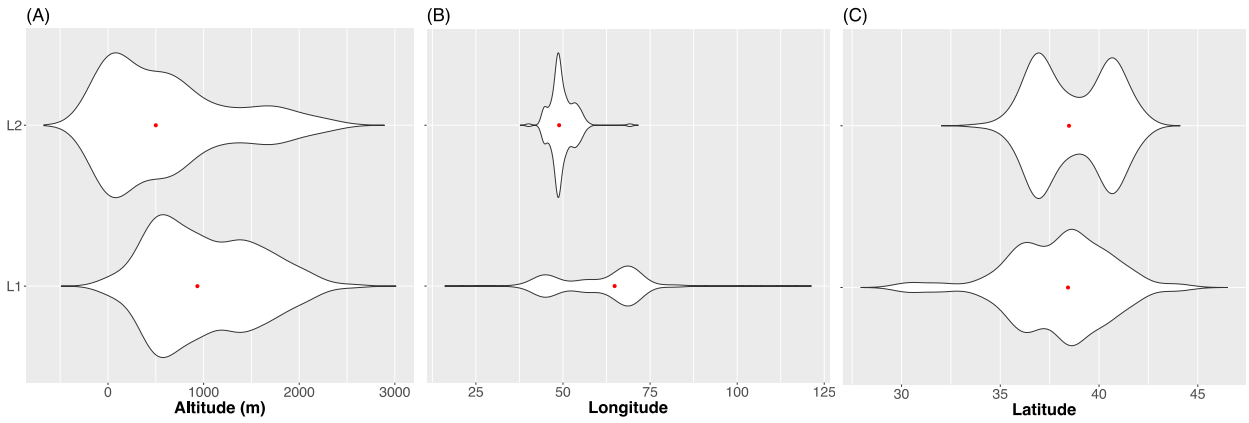

Figure C.10. Violin plots showing L1 and L2 distribution for altitude, longitude and latitude.

Red dots are median values.

Appendix D - Supplementary material Chapter 4

This appendix contains supplementary figures and tables for Chapter 4.

Table D.1. Pedigree information of mapping populations.

Mapped gene	***Population code***	***Pedigree***	***Final population size***	***Reported in***
H5	H5-EN	'Newton'/Erin	146	(Patterson *et al.* 1994)
H10	H10-JN	'Newton'/Joy	150	(Patterson *et al.* 1994)
H13	H13-MN	'Newton'/Molly	113	(Patterson *et al.* 1994)
H13	H13-MO	'Overley'/Molly	87	This study
H2147	Fam1	'Overley'*4/*KU2147*	316	This study
H2147	Fam2	'Overley'*4/*KU2147*	273	This study

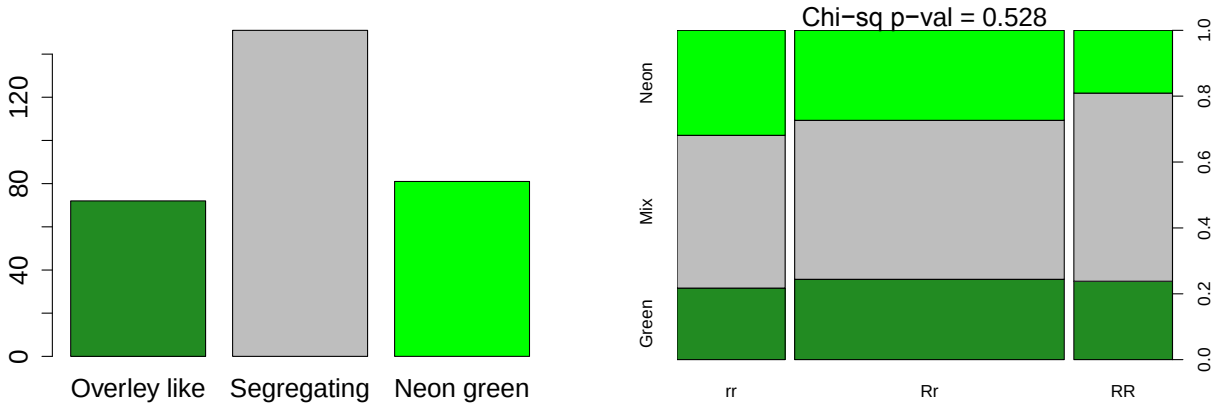

Figure D.1. Chi-square test of association between plant color and Hessian fly resistance.

(Left panel) Barplot showing distribution of different plant color classes in Fam1. (Right panel) Stacked barplot showing distribution of colors within different genotypic classes. Chi-square *p*-value is shown at the top of the plot.

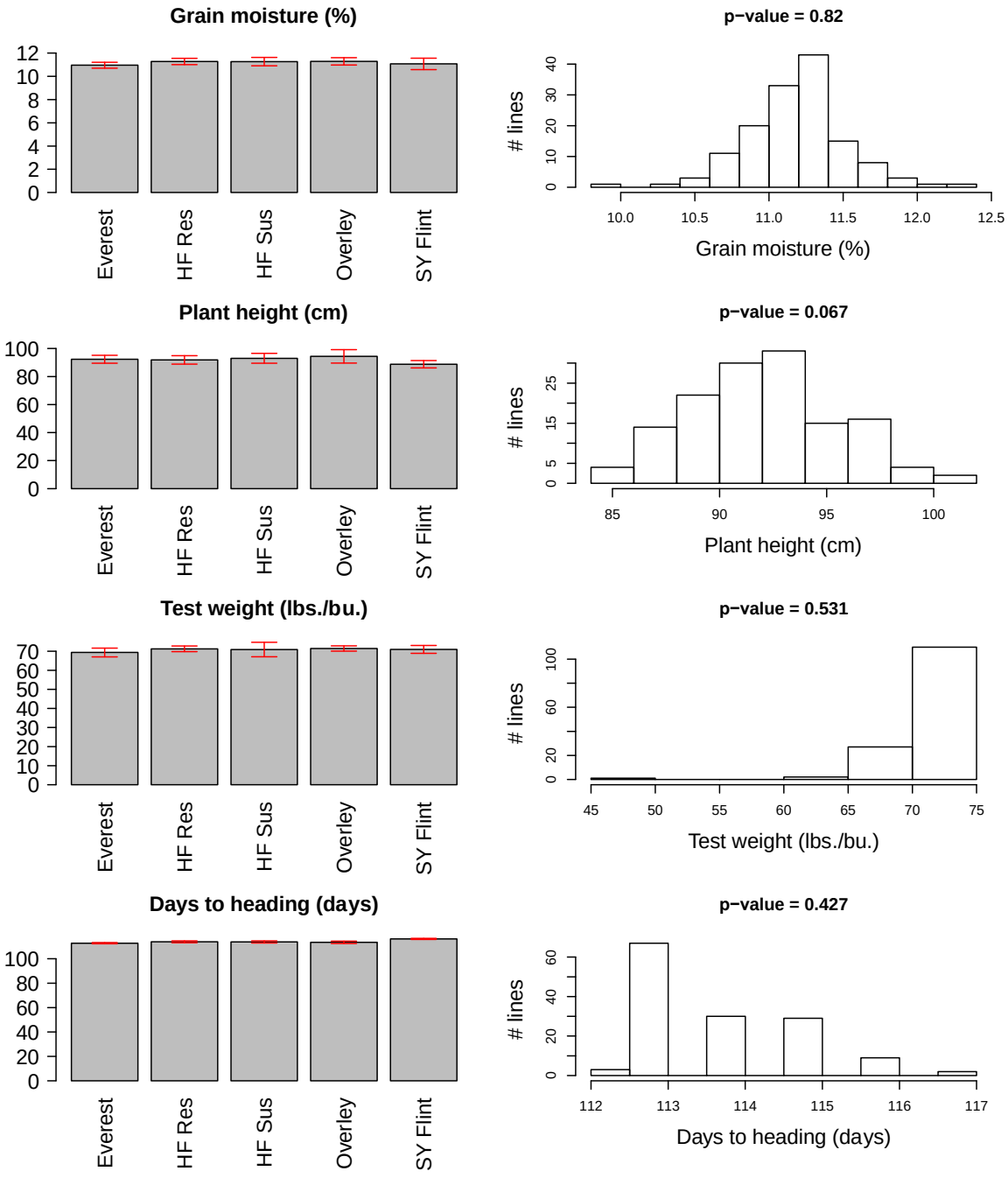

Figure D.2. Phenotypic data for grain moisture, plant height, test weight, days to heading.

(Left panel) Barplots showing the distribution of grain moisture, plant height, test weight, days to heading for two controls, 'Everest' and 'SY Flint', resistant lines, susceptible lines, and recurrent parent 'Overley'. (Right panel) Distribution of these traits for the whole population. Welch t-test *p*-value is shown on the top of histogram.