DEVELOPMENT OF GENOMIC RESOURCES TO FACILITATE PLANT BREEDING

By

Nolan Bornowski

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Breeding, Genetics and Biotechnology – Plant Biology – Doctor of Philosophy

2022

PUBLIC ABSTRACT

DEVELOPMENT OF GENOMIC RESOURCES TO FACILITATE PLANT BREEDING

By

Nolan Bornowski

In the past decade DNA sequencing has become more affordable and computers have become more powerful. These technological developments resulted in an explosion of research on plant genomics, the study of the DNA content of plants. As a result, genomics-based approaches are widely-adopted by modern plant breeders with the desire to improve plants in response to the food security and climate change challenges in the coming decades. This dissertation describes the generation of genomic resources for several important plant types: culinary herbs belonging to the mint family, maize (corn) inbred lines from the stiff stalk heterotic group, and tepary beans, which are more highly heat- and drought-resistant compared to their sister species, common bean. These research projects provide plant breeders the tools to assess genetic diversity and ultimately make better plants for humankind.

ABSTRACT

DEVELOPMENT OF GENOMIC RESOURCES TO FACILITATE PLANT BREEDING

By

Nolan Bornowski

Recent advances in sequencing and computation power have greatly contributed to our knowledge of plant genomics, and the development and use of plant genomic resources will be critical as plant researchers and breeders address future food security in light of the increasing world population, decreasing arable land, and variable effects of climate change. Plants belonging to the mint family provide culinary, medicinal, and cultural value due to their production of secondary metabolites. Genome assemblies and annotations for four important culinary herbs were generated to highlight genes involved in terpenoid biosynthetic pathways. Maize (*Zea mays* L.) is the most produced crop worldwide due in part to extensive commercial breeding programs. Genome assemblies and annotations for five commercially relevant maize inbred lines belonging to the stiff-stalk heterotic group were generated to characterize the pan-stiff-stalk gene repertoire and genomic regions associated with these founder lines. Tepary bean (*Phaseolus acutifolius* A. Gray), a close relative of the common bean (*Phaseolus vulgaris* L.), is indigenous to the arid climates of northern Mexico and produces high seed yields under drought stress. A diverse panel of tepary bean accessions was assembled, genotyped, and phenotyped to identify genomic regions associated with key agronomic traits that can be harnessed for tepary bean improvement.

This dissertation is dedicated to the late John C. “Jack” Bogle for his pioneering work with index funds.

ACKNOWLEDGEMENTS

I want to thank my advisor Dr. Robin Buell for introducing me to the field of plant genomics. She has challenged me to continually improve myself as a person and develop into an independent researcher. It has been an honor to be a part of her lab group and witness how she has been a stalwart mentor and an inspiration for myself and many other young scientists. I want to acknowledge the strong comradery with my fellow plant science graduate students, with whom I have shared classes, ideas, laughs, and life moments. I am obligated to acknowledge my family, who have enthusiastically supported me despite not fully understanding my research. I am appreciative of Lansing and all of its rustic charm, and for the time spent exploring its parks, activities, and establishments, especially Dagwood's Tavern. Finally, I cannot express in words how much the love and support of my friend and partner, Amber Bassett, has meant to me these past years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

OVERVIEW

The discipline of plant genomics involves the study of the genomes of plant species. From a plant biologist's perspective, the "plant" in plant genomics has typically been a model organism like Chlamydomonas or Arabidopsis, whereas from a plant breeder's perspective, the "plant" in plant genomics is often a crop plant with food security and/or economic importance like beans or maize. As a graduate student belonging to both the Department of Plant Biology and the interdisciplinary program of Plant Breeding, Genetics, and Biotechnology, I have studied plant genomics at the intersection of these foci. Recent technological advances in sequencing and computing have made the study of plant genomics more accessible to plant scientists, including breeders, regardless of their organism of focus, which has been a driving motivation for my dissertation research.

GENOMICS

Genomics is the field of study concerning the complete genome content of an organism including the nucleus and any organellar genomes. Characterizing the genomic features such as the genes, repetitive elements, and other non-coding material is a key objective in genomic studies. Genomic features can be analyzed for a single individual, a group of organisms (population genomics), or even temporally across species or clades (evolutionary genomics). As

the underlying DNA sequence of a genome is central to all biological principles, a complete understanding of genomes is fundamental to enabling new biological discoveries.

SEQUENCING

Obtaining genomic DNA sequence has been possible since the late 1970's. These so-called first-generation DNA sequencing methods relied on the incorporation of a terminal, radioactively-labeled phosphorous-32 to a DNA fragment. According to the chemical sequencing method of Maxam-Gilbert, the DNA fragment was then cleaved at either G, C, A+G, or C+G sites and then size-separated on an acrylamide gel. The presence or absence of the various radiolabeled fragments were used to decipher the original DNA fragment sequence (Maxam and Gilbert, 1977). A more widely-used method was developed by Frederick Sanger, known as the chain-termination sequencing method, or Sanger sequencing. This method works by using a DNA polymerase enzyme to incorporate deoxyribonucleotides and a smaller proportion of modified deoxyribonucleotides called dideoxyribonucleotides (ddNTPs) to a growing DNA strand using a primer and a DNA template. The extension of the sequence by the DNA polymerase is halted by the random addition of the ddNTP, resulting in DNA fragments of varying length. These DNA fragments are then separated based on length using gel electrophoresis, with the smaller fragments travelling at a faster rate than larger fragments (Sanger et al., 1977). Initially, Sanger sequencing used radioactivity to detect the synthesized fragments but this was later replaced with fluorescently labeled ddNTPs. Sanger sequencing remains useful for determining the sequence of small DNA fragments (< 800 nucleotides), but sequencing large genomes using whole genomic DNA and Sanger sequencing is impractical.

A revolution in genomics occurred in 1995 as a new method was introduced called whole genome shotgun (WGS) sequencing that enabled generation of the complete genome sequencing of living organisms. This method involves fragmenting high-quality genomic DNA, cloning the fragments into a plasmid, isolating plasmid DNA, sequencing the ends of the insert using Sanger sequencing, and then assembling the genome from these 'shotgun reads.' This was done first with *Haemophilus influenzae* in 1995, a 1,830,137 bp bacterial genome (Fleischmann et al., 1995). This accomplishment also accelerated the development of high-capacity Sanger sequencing machines by Applied Biosystems capable of sequencing 96 reactions at a time generating read lengths of 700-800 nucleotides.

In the mid 2000s, revolutionarily new sequencing technologies termed 'next-generation sequencing' emerged that employed a sequencing-by-synthesis approach in which DNA fragments were directly sequenced on a flowcell with each new nucleotide detected via fluorescently-tagged nucleotides (reviewed in (Shendure and Ji, 2008; Mardis, 2013). This sequencing approach was able to efficiently and economically generate millions of short (~100 nucleotide) sequences ("reads") and initially ~30 nucleotides that were used in whole genome assembly. More recently, advances in sequencing technologies and computational resources have ushered in a new era of genomics with 'third generation sequencing methods' that can generate much longer reads (>5,000 to 100,000 nucleotides) and solve outstanding challenges in genomics such as centromere sequencing (Rabanal et al., 2022) and telomere-to-telomere assembly (Nurk et al., 2022). Two major companies have emerged in the field of long read sequencing, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio long reads are generated by ligating hairpin adapters to the ends of a double-stranded template DNA fragment,

and then having a DNA polymerase synthetize a complimentary strand with fluorescently-labeled nucleotides. As each nucleotide is incorporated, a unique, small light signal is emitted. These pulses of light are recorded like a movie, and when the "movie" is replayed, the sequence of the DNA fragment can be determined from the signals (Rhoads and Au, 2015). One of the major advantages of this method is that a single DNA molecule can be sequenced multiple times as it passes through the DNA polymerase. These "sub-reads" can be combined to form a highly accurate circular consensus sequence ("CCS"), which provides a more accurate representation of the target DNA (Travers et al., 2010). Meanwhile, ONT long reads are generated by guiding DNA fragments through an transmembrane protein pore and measuring the fluctuations in electrical conductance as the nucleotides pass through the protein (reviewed in (Wang et al., 2021). A motor protein anchored above the pore acts as a helicase to ratchet a single strand at a time and control the rate of movement through the pore. The underlying sequence of the DNA fragment can be inferred ("base-called") in real-time or after a sequencing run by computational algorithms that deconvolute the electrical signal into nucleotides. Regardless of which third generation technology is used, access to long reads facilitates the assembly of repetitive genome features such as telomeres, centromeres, tandem repeats, and transposable elements, which can create a more complete representation of the actual chromosome during genome assembly.

Genome assembly effort was possible due to improvements in computers (hardware) and algorithms (software) to assemble the whole genome shotgun reads. In brief, reads with overlapping sequences are joined together to form larger sequences called contigs, and the contigs are then oriented and assembled into even larger sequences called scaffolds. The scaffolds were then be concatenated into even larger sequences called pseudomolecules that

represent individual chromosomes. The combination of pseudomolecules and scaffolds is considered the nuclear genome assembly.

PLANT GENOMICS AND APPLICATIONS FOR BREEDING

The simultaneous advances in sequencing throughput and computational power resulted in a series of genome assembly milestones. The model plant *Arabidopsis thaliana* (L.) Heynh. was the first plant genome to be assembled (Arabidopsis Genome Initiative, 2000). Already widely utilized by plant biologists for its compact architecture, diploid chromosome content, rapid life cycle, fecundity, and ease of transformation, the *A. thaliana* genome is small compared to other plants (~157 Mbp) with few repetitive elements. Throughout the two decades since its release, the *A. thaliana* genome and gene annotations have been revised and are widely used as a basis for plant genomic studies. Though extremely useful as a model plant, *A. thaliana* is not a crop. The first crop plant genome to be assembled was rice (*Oryza sativa* var. *japonica*), which also had a relatively small genome (~390 Mbp) (International Rice Genome Sequencing Project, 2005), and could be used in translational research across other important cereal crops like maize and wheat (Jackson, 2016). Other crop plant genome assemblies soon followed, such as grape (Jaillon et al., 2007), sorghum (Paterson et al., 2009), maize (Schnable et al., 2009), soybean (Schmutz et al., 2010), potato (Potato Genome Sequencing Consortium et al., 2011), and tomato (Tomato Genome Consortium, 2012). These genome assemblies and annotations ushered in a new era of crop improvement- facilitating orthologous gene identification, molecular marker development, and trait mapping studies (Figure 1.1). As of 2022, there are more than 800 publicly available plant genome assemblies (Marks et al., 2021), however that is still just a small fraction of the estimated ~400,000 extant plant species (Enquist et al., 2019). Clearly, more

genome assembly and annotation projects are forthcoming, and these resources can help plant breeders and researchers better understand plant diversity not only in the context of DNA content, but also regarding transcriptomic, epigenomic, and metabolomic potential for improving plants.

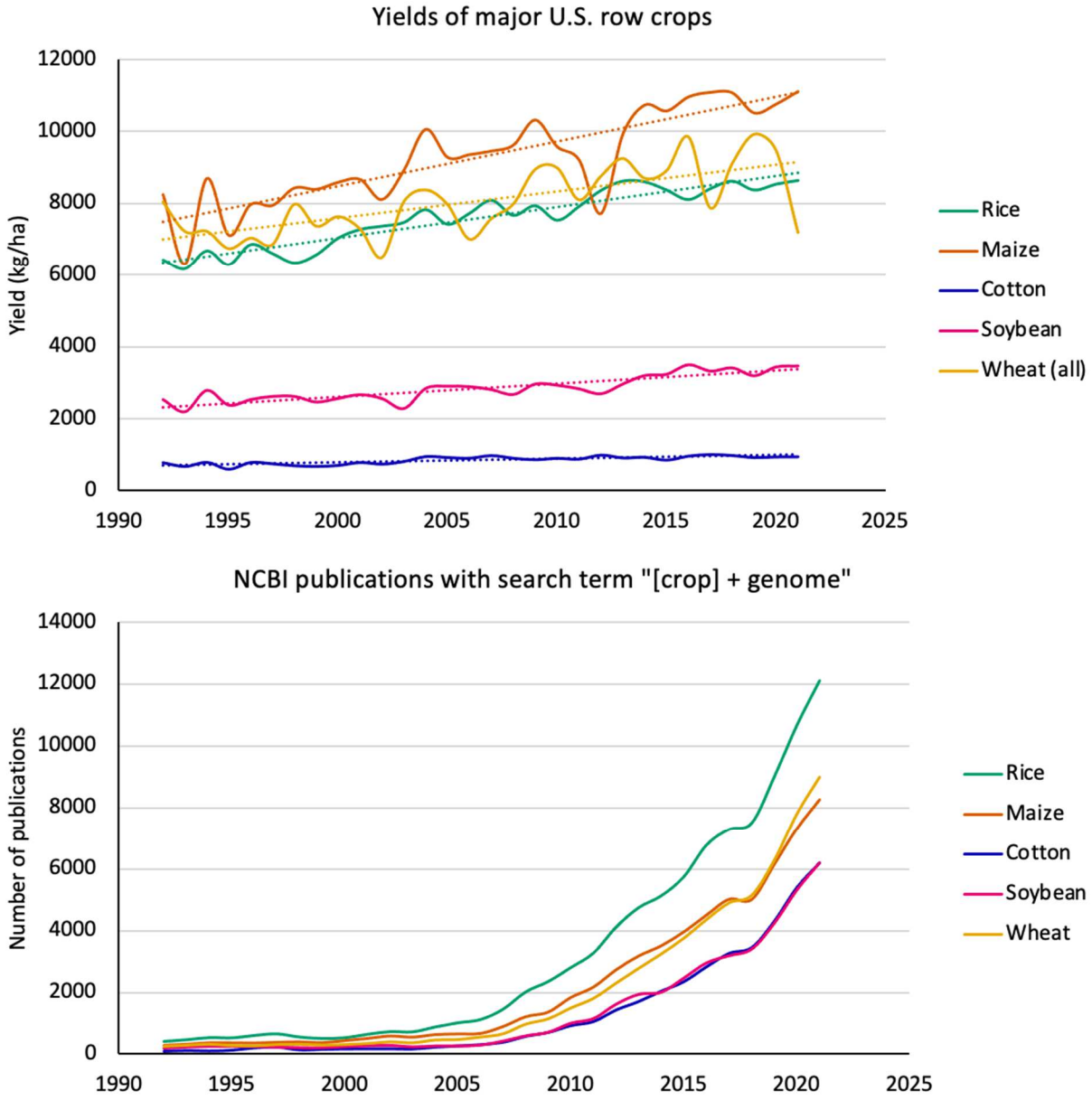


Figure 1.1. Annual yields and publication counts of US major row crops.

Yield data was retrieved from USDA-NASS accessed from https://www.nass.usda.gov/Charts_and_Maps/Field_Crops/index.php. Publication counts were

Figure 1.1 (cont'd)

obtained from querying the common name of the crop and "genome" in the NCBI PMC query accessed from https://www.ncbi.nlm.nih.gov/pmc/.

Mint Genomics and Breeding Resources

The mint family of plants, Lamiaceae, contains a staggering number of not only species diversity, but also metabolite diversity (Weng et al., 2012; Lange, 2015). Within the Lamiaceae, the largest subfamily, Nepetoideae, contains approximately half of all mint species, and the greatest amount of monoterpene diversity compared to the other mint subfamilies (El-Gazzar and Watson, 1970). Many of these Nepetoideae species have culinary, medicinal, ornamental, cultural, and ecological importance including lavender, salvia, peppermint, spearmint, and catmint. Several species within Nepetoideae are widely used for their culinary properties because they bestow unique flavor profiles to dishes when added as ingredients or flavorings. These flavors and aromas are largely due to their production of a class of secondary metabolites called terpenoids. Terpenoids are synthesized by enzymes called terpene synthases (TPS) and can be subsequently decorated via modifying enzymes such as CYP450s and UGTs. In the publication for Chapter 2 (Bornowski et al., 2020a), I present the genome assembly, annotation, and comparative genomics for four culinary herbs- Sweet basil (*Ocimum basilicum* L.), Oregano (*Origanum vulgare* L.), Sweet marjoram (*Origanum majorana* L.), and Rosemary (*Rosmarinus officinalis* L.)- and highlight their terpenoid profiles and repertoire of terpenoid biosynthetic genes. Recent phylogenomic research on the Nepetoideae uncovered not only the evolutionary relationships between Nepetoideae members, but also quantified their secondary metabolite profiles and characterized genes involved in the evolution of the underlying biosynthetic

pathways leading to terpene diversity in the Lamiaceae (Boachon et al., 2018). Generation of genomic resources of key species in this important mint clade can facilitate more informed breeding decisions with regards to developing chemotypes containing unique metabolite profiles (Rodríguez-Solana et al., 2014) and identifying orthologous genes and their alleles that synthesize or modify secondary metabolites (Weng and Noel, 2012).

Maize Genomics and Breeding Resources

Maize (*Zea mays* L.) is the most produced cereal in the world on a per-weight basis, and is projected to increase in acreage in the coming decades (Erenstein et al., 2021). Its current importance as an economic and food security crop is reflected by the extent of breeding efforts undertaken. Originating as a weedy, small-grained grass from Mexico (teosinte), maize has been domesticated and dramatically improved upon over the past 7,000 years (Goodman and Galinat, 1988). Commercial maize has progressed from open-pollinated varieties to double- and single-cross hybrids, to the present day hybrids that can be modeled from genome predictions (Technow et al., 2014) and contain genetically engineered resistances to biotic (Koch et al., 2015) and abiotic stressors (Adee et al., 2016). The maize genome was released in 2009 from the reference genotype, B73 (Schnable et al., 2009). Since then, the genome and annotation have become more complete as new sequencing technologies matured permitting assembly and annotation of numerous maize inbreds across different germplasm pools (so-called "heterotic groups") (Duvick, 2005) The assembly and annotation of multiple maize lines is particularly important because substantial intra-species breeding has occurred within the heterotic groups (Mikel, 2011), and extreme presence-absence variation has been documented (Springer et al., 2009).

Several maize diversity panels have been generated to consolidate the extreme genetic diversity of modern maize into a representative germplasm set for quantitative genetic study. The Wisconsin Diversity Panel (WiDiv) contains 627 inbred lines adapted to Wisconsin and northern latitudes (Hansey et al., 2011), although a subset of 60 lines within this diversity panel was found to contain 90% of the haplotype diversity in the larger panel (Yan et al., 2009). The WiDiv has been used in numerous genetic mapping studies, including virus resistance (Gage et al., 2019) and kernel composition traits (Renk et al., 2021), and was later expanded by (Mazaheri et al., 2019) to 942 inbred lines for mapping stalk agronomic traits. The WiDiv and other accessions were also used to characterize the maize pan-transcriptome (Hirsch et al., 2014) and transposable element diversity (Qiu et al., 2021). Another maize diversity panel, the called the Ames Diversity Panel, is also widely used for quantitative mapping studies, and contains approximately 2500 inbred lines. Since initial genotyping was conducted (Romay et al., 2013), inbreds from this panel have been used in studies mapping root development (Pace et al., 2015), grain metabolites (Wu et al., 2021), flowering time (Li et al., 2016), and pan-genome anchors (Lu et al., 2015). Various other maize diversity panels have been assembled and used for genomic prediction (Windhausen et al., 2012; Gowda et al., 2015; Rio et al., 2019; Allier et al., 2020) or to shed light on evolution and domestication (Tian et al., 2009; Wang et al., 2017). The genotypic and phenotypic characterization of the aforementioned maize panels has greatly facilitated quantitative genetics and provided insight on the substantial diversity found in maize.

Maize lines belonging to the stiff stalk heterotic group were initially derived from a synthetic population called Iowa Stiff-Stalk Synthetic that had low prevalence of lodging and high yields (Troyer, 2004, 1999). Stiff stalk germplasm has been widely used in commercial

hybrid production, whereby crossing a stiff stalk maize inbred line with a non-stiff stalk maize inbred line generated substantial heterosis (i.e. yield and vigor) in the F1 hybrid progeny. The inbred line combinations that produced the best hybrids were legally protected from use by other companies through Plant Variety Protection (PVP) certifications. PVP certifications give the organization sole ownership of the germplasm for 20 years, after which it can be freely used by other organizations (White et al., 2020). In the publication for Chapter 3 (Bornowski et al., 2021), I present the genome assembly, annotation, and comparative genomics for five commercially relevant maize stiff stalk inbreds with the reference genotype of the B73 stiff stalk inbred, which serves as a valuable resource to better understand the genomic regions targeted directly or indirectly by commercial breeding institutions.

Tepary Bean Genomics and Breeding Resources

The *Phaseolus* plant lineage encompasses plants that are cultivated around the world as a human food source. While five species have been domesticated, the common bean (*Phaseolus vulgaris* L.) is the most consumed food legume globally (Broughton et al., 2003), and this is also reflected in the breeding progress and genomic resources available to breeders. In the decades prior to the common bean genome assembly, bean breeders were quick to adapt gel-based molecular markers to select for critical disease resistance traits (Kelly and Bornowski, 2018). A series of DNA microarrays ("SNP chips") were developed from polymorphisms across a draft assembly and common bean varieties representing different market classes (Song et al., 2015). These microarrays made genotyping more accessible to breeders, and were subsequently utilized to genotype hundreds of common bean lines in diversity panels such as the Andean Diversity Panel (ADP) (Cichy et al., 2015) and the Middle American Diversity Panel (MDP) (Moghaddam

et al., 2016). The microarray genotyping of these diversity panels opened the door for a flurry of genetic studies that is ongoing to the present-day. With genotypes in-hand, researchers were able to concentrate their efforts toward phenotypic screening of the panels for traits of interest. This had led to quantitative mapping studies on agronomic traits (Kamfwa et al., 2015; Moghaddam et al., 2016), flooding tolerance (Soltani et al., 2018), anthracnose resistance (Zuiderveen et al., 2016), and cooking quality traits (Bassett et al., 2021; Katuuramu et al., 2018), among others. In 2015, the first common bean genome assembly was published (Schmutz et al., 2014), and the authors demonstrated high collinearity across common bean and its relative, soybean (*Glycine max* L.). This finding was significant because it allowed common bean researchers to transitively harness the more-developed soybean genomic resources to fill in knowledge gaps in common bean genomics. In the following years, additional genome assemblies have been published for common bean: OAC Rex (https://www.ncbi.nlm.nih.gov/genome/380?genome_assembly_id=1500596), breeding line BAT93 (https://www.ncbi.nlm.nih.gov/genome/380?genome_assembly_id=262776, (Vlasova et al., 2016; Rendón-Anaya et al., 2017) cultivar Pinto UI111 (https://phytozome-next.jgi.doe.gov/info/PvulgarisUI111_v1_1) cultivar Labor Ovalle (https://phytozome-next.jgi.doe.gov/info/PvulgarisLaborOvalle_v1_1), and other related species like cowpea (*Vigna unguiculata* [L.] Walp.) (Lonardi et al., 2019), lima bean (*Phaseolus lunatus* L.) (Garcia et al., 2021), and tepary bean (*Phaseolus acutifolius* A. Gray) (Moghaddam et al., 2021). These genomic resources will be a boon for breeders and researchers working to improve leguminous crops.

Tepary bean (*P. acutifolius*) is a prime example of a *Phaseolus* species that has undergone minimal breeding progress that can now be accelerated by the development of genomic resources. The recent publication of a tepary reference genome and annotation highlighted its ability to withstand extreme water scarcity conditions and genic conservation across *Phaselous* spp (Moghaddam et al., 2021). These genomic resources can be leveraged to better understand diversity and breeding targets within tepary by screening an assembled panel of tepary accessions for genetic diversity, as has been done with great success in common bean. The tepary bean diversity panel (TDP) consists of 423 tepary accessions and includes cultivated, wild, and weedy tepary lines that were phenotyped for key traits and genotyped for single nucleotide polymorphisms (SNPs) relative to the tepary bene reference genome cultivar, Frijol Bayo. In Chapter 4, I present the statistical association of trait measurements and SNPs using a genome-wide association study (GWAS) to identify genomic regions underlying traits critical for tepary bean breeding and improvement.

DISSERTATION PROJECTS AND SIGNIFICANCE

The second chapter of this dissertation presents the genome assembly, annotation, and comparative genomics of four culinary herbs within the Nepetoideae subfamily of mints (Lamiaceae), and terpene synthases and related biosynthetic enzymes in secondary metabolism were characterized. The third chapter of this dissertation presents the genome assembly, annotation, and comparative genomics of five commercially important maize inbred lines belonging to the stiff stalk heterotic group and comparison with the reference B73 assembly, another member of the stiff stalk heterotic group. A re-annotation of the reference genome variety, B73, was generated using only cognate transcript evidence to prevent technical artefacts

during intra-heterotic group comparisons. Furthermore, genomic regions passed down by these founding lines were identified, revealing important haplotypes that have been selected for and conserved during commercial maize breeding. The fourth chapter of this dissertation presents a quantitative genetic analysis of a diverse panel of tepary beans to reveal loci associated with key agronomic and morphological traits that can be used to further tepary breeding and genetic improvement.

REFERENCES

REFERENCES

Adee, E., Roozeboom, K., Balboa, G.R., Schlegel, A., and Ciampitti, I.A. (2016). Drought-tolerant corn hybrids yield more in drought-stressed environments with no penalty in non-stressed environments. Front. Plant Sci. **7**: 1534.

Allier, A., Teyssèdre, S., Lehermeier, C., Charcosset, A., and Moreau, L. (2020). Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. Züchter Genet. Breed. Res. **133**: 201–215.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**: 796–815.

Bassett, A., Kamfwa, K., Ambachew, D., and Cichy, K. (2021). Genetic variability and genome-wide association analysis of flavor and texture in cooked beans (*Phaseolus vulgaris* L.). Züchter Genet. Breed. Res. **134**: 959–978.

Boachon B, Buell CR, Crisovan E, Dudareva N, Garcia N, Godden G, Henry L, Kamileen MO, Kates HR, Kilgore MB, et al (2018) Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. Mol Plant **11**: 1084–1096

Bornowski N, Michel KJ, Hamilton JP, Ou S, Seetharam AS, Jenkins J, Grimwood J, Plott C, Shu S, Talag J, et al (2021) Genomic variation within the maize stiff-stalk heterotic germplasm pool. Plant Genome **14**: e20114

Bornowski, N., Hamilton, J.P., Liao, P., Wood, J.C., Dudareva, N., and Buell, C.R. (2020). Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae. DNA Res. **27**.

Broughton, W.J., Hernández, G., Blair, M., Beebe, S., Gepts, P., and Vanderleyden, J. (2003). Beans (*Phaseolus* spp.) – model food legumes. Plant Soil **252**: 55–128.

Cichy KA, Porch TG, Beaver JS, Cregan P, Fourie D, Glahn RP, Grusak MA, Kamfwa K, Katuuramu DN, McClean P, et al (2015) A *Phaseolus vulgaris* Diversity Panel for Andean Bean Improvement. Crop Sci **55**: 2149–2160

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

El-Gazzar, A. and Watson, L. (1970). A taxonomic study of labiatae and related genera. New Phytol. **69**: 451–486.

Enquist BJ, Feng X, Boyle B, Maitner B, Newman EA, Jørgensen PM, Roehrdanz PR, Thiers BM, Burger JR, Corlett RT, et al (2019) The commonness of rarity: Global and future distribution of rarity across land plants. Sci Adv **5**: eaaz0414

Erenstein, O., Chamberlin, J., and Sonder, K. (2021). Estimating the global number and distribution of maize and wheat farms. Glob. Food Sec. **30**: 100558.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**: 496–512.

Gage, J.L., Vaillancourt, B., Hamilton, J.P., Manrique-Carpintero, N.C., Gustafson, T.J., Barry, K., Lipzen, A., Tracy, W.F., Mikel, M.A., Kaeppler, S.M., Buell, C.R., and de Leon, N. (2019). Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. Plant Genome **12**: 180069.

Garcia T, Duitama J, Zullo SS, Gil J, Ariani A, Dohle S, Palkovic A, Skeen P, Bermudez-Santana CI, Debouck DG, et al (2021) Comprehensive genomic resources related to domestication and crop improvement traits in Lima bean. Nat Commun **12**: 702

Goodman, M.M. and Galinat, W.C. (1988). The history and evolution of Maize. CRC Crit. Rev. Plant Sci. **7**: 197–220.

Gowda, M., Das, B., Makumbi, D., Babu, R., Semagn, K., Mahuku, G., Olsen, M.S., Bright, J.M., Beyene, Y., and Prasanna, B.M. (2015). Genome-wide association and genomic prediction of resistance to maize lethal necrosis disease in tropical maize germplasm. Züchter Genet. Breed. Res. **128**: 1957–1968.

Hansey, C.N., Johnson, J.M., Sekhon, R.S., Kaeppler, S.M., and Leon, N. (2011). Genetic diversity of a maize association population with restricted phenology. Crop Sci. **51**: 704–715.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell **26**: 121–135

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature **436**: 793–800.

Jackson, S.A. (2016). Rice: The first crop genome. Rice (N. Y.) **9**: 14.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**: 463–467

Kamfwa, K., Cichy, K.A., and Kelly, J.D. (2015). Genome-wide association analysis of symbiotic nitrogen fixation in common bean. Züchter Genet. Breed. Res. **128**: 1999–2017.

Katuuramu, D.N., Hart, J.P., Porch, T.G., Grusak, M.A., Glahn, R.P., and Cichy, K.A. (2018). Genome-wide association analysis of nutritional composition-related traits and iron bioavailability in cooked dry beans (*Phaseolus vulgaris* L.). Mol. Breed. **38**.

Kelly, J.D. and Bornowski, N. (2018). Marker-assisted breeding for economic traits in common bean. In Biotechnologies of Crop Improvement, Volume 3 (Springer International Publishing: Cham), pp. 211–238.

Koch, M.S., Ward, J.M., Levine, S.L., Baum, J.A., Vicini, J.L., and Hammond, B.G. (2015). The food and environmental safety of Bt crops. Front. Plant Sci. **6**: 283.

Lange, B.M. (2015). The evolution of plant secretory structures and emergence of terpenoid chemical diversity. Annu. Rev. Plant Biol. **66**: 139–159.

Li Y-X, Li C, Bradbury PJ, Liu X, Lu F, Romay CM, Glaubitz JC, Wu X, Peng B, Shi Y, et al (2016) Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. Plant J **86**: 391–402

Lonardi S, Muñoz-Amatriaín M, Liang Q, Shu S, Wanamaker SI, Lo S, Tanskanen J, Schulman AH, Zhu T, Luo M-C, et al (2019) The genome of cowpea (*Vigna unguiculata* [L.] Walp.). Plant J **98**: 767–782

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, et al (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun **6**: 6914

Mardis, E.R. (2013). Next-generation sequencing platforms. Annu. Rev. Anal. Chem. (Palo Alto Calif.) **6**: 287–303.

Marks, R.A., Hotaling, S., Frandsen, P.B., and VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. Nat. Plants **7**: 1571–1578.

Maxam, A.M. and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. **74**: 560–564.

Mazaheri M, Heckwolf M, Vaillancourt B, Gage JL, Burdo B, Heckwolf S, Barry K, Lipzen A, Ribeiro CB, Kono TJY, et al (2019) Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol **19**: 45

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

Moghaddam SM, Mamidi S, Osorno JM, Lee R, Brick M, Kelly J, Miklas P, Urrea C, Song Q, Cregan P, et al (2016) Genome-wide association study identifies candidate loci underlying agronomic traits in a Middle American Diversity Panel of common bean. Plant Genome. doi: 10.3835/plantgenome2016.02.0012

Moghaddam SM, Oladzad A, Koh C, Ramsay L, Hart JP, Mamidi S, Hoopes G, Sreedasyam A, Wiersma A, Zhao D, et al (2021) The tepary bean genome provides insight into evolution and domestication under heat stress. Nat Commun **12**: 2638

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al (2022) The complete sequence of a human genome. Science **376**: 44–53

Pace, J., Gardner, C., Romay, C., Ganapathysubramanian, B., and Lübberstedt, T. (2015). Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). BMC Genomics **16**: 47.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**: 551–556

Potato Genome Sequencing Consortium, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, et al (2011) Genome sequence and analysis of the tuber crop potato. Nature **475**: 189–195

Qiu, Y., O'Connor, C.H., Della Coletta, R., Renk, J.S., Monnahan, P.J., Noshay, J.M., Liang, Z., Gilbert, A., Anderson, S.N., McGaugh, S.E., Springer, N.M., and Hirsch, C.N. (2021). Whole-genome variation of transposable element insertions in a maize diversity panel. G3 (Bethesda) **11**.

Rabanal, F.A., Gräff, M., Lanz, C., Fritschi, K., Llaca, V., Lang, M., Carbonell-Bejerano, P., Henderson, I., and Weigel, D. (2022). Pushing the limits of HiFi assemblies reveals centromere diversity between two *Arabidopsis thaliana* genomes. bioRxiv: 2022.02.15.480579.

Rendón-Anaya M, Montero-Vargas JM, Saburido-Álvarez S, Vlasova A, Capella-Gutierrez S, Ordaz-Ortiz JJ, Aguilar OM, Vianello-Brondani RP, Santalla M, Delaye L, et al (2017) Genomic history of the origin and domestication of common bean unveils its closest sister species. Genome Biol. doi: 10.1186/s13059-017-1190-6

Renk, J.S., Gilbert, A.M., Hattery, T.J., O'Connor, C.H., Monnahan, P.J., Anderson, N., Waters, A.J., Eickholt, D.P., Flint-Garcia, S.A., Yandeau-Nelson, M.D., and Hirsch, C.N. (2021). Genetic control of kernel compositional variation in a maize diversity panel. Plant Genome **14**: e20115.

Rhoads, A. and Au, K.F. (2015). PacBio Sequencing and its applications. Genomics Proteomics Bioinformatics **13**: 278–289.

Rio, S., Mary-Huard, T., Moreau, L., and Charcosset, A. (2019). Genomic selection efficiency and a priori estimation of accuracy in a structured dent maize panel. Züchter Genet. Breed. Res. **132**: 81–96.

Rodríguez-Solana, R., Daferera, D.J., Mitsi, C., Trigas, P., Polissiou, M., and Tarantilis, P.A. (2014). Comparative chemotype determination of Lamiaceae plants by means of GC–MS, FT-IR, and dispersive-Raman spectroscopic techniques and GC-FID quantification. Ind. Crops Prod. **62**: 22–33.

Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, et al (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol **14**: R55

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**: 5463–5467.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature **463**: 178–183

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet **46**: 707–713

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326**: 1112–1115

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. **26**: 1135–1145.

Soltani A, MafiMoghaddam S, Oladzad-Abbasabadi A, Walter K, Kearns PJ, Vasquez-Guzman J, Mamidi S, Lee R, Shade AL, Jacobs JL, et al (2018) Genetic analysis of flooding tolerance in an Andean diversity panel of dry bean (*Phaseolus vulgaris* L.). Front Plant Sci. doi: 10.3389/fpls.2018.00767

Song, Q., Jia, G., Hyten, D.L., Jenkins, J., Hwang, E.-Y., Schroeder, S.G., Osorno, J.M., Schmutz, J., Jackson, S.A., McClean, P.E., and Cregan, P.B. (2015). SNP assay development for linkage map construction, anchoring whole-genome sequence, and other genetic and genomic applications in common bean. G3 (Bethesda) **5**: 2285–2290.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet **5**: e1000734

Technow, F., Schrag, T.A., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A.E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics **197**: 1343–1355.

Tian, F., Stevens, N.M., and Buckler, E.S., 4th (2009). Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc. Natl. Acad. Sci. U. S. A. **106 Suppl 1**: 9979–9986.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635–641.

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., and Turner, S.W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. **38**: e159.

Troyer, A.F. (1999). Background of U.S. hybrid corn. Crop Sci. **39**: 601–626.

Troyer, A.F. (2004). Background of U.S. hybrid corn II. Crop Sci. **44**: 370–380.

Vlasova A, Capella-Gutiérrez S, Rendón-Anaya M, Hernández-Oñate M, Minoche AE, Erb I, Câmara F, Prieto-Barja P, Corvelo A, Sanseverino W, et al (2016) Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. Genome Biol **17**: 32

Wang, L., Beissinger, T.M., Lorant, A., Ross-Ibarra, C., Ross-Ibarra, J., and Hufford, M.B. (2017). The interplay of demography and selection during maize domestication and expansion. Genome Biol. **18**: 215.

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K.F. (2021). Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol. **39**: 1348–1365.

Weng, J.-K. and Noel, J.P. (2012). The remarkable pliability and promiscuity of specialized metabolism. Cold Spring Harb. Symp. Quant. Biol. **77**: 309–320.

Weng, J.-K., Philippe, R.N., and Noel, J.P. (2012). The rise of chemodiversity in plants. Science **336**: 1667–1670.

White, M.R., Mikel, M.A., Leon, N., and Kaeppler, S.M. (2020). Diversity and heterotic patterns in North American proprietary dent maize germplasm. Crop Sci. **60**: 100–114.

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J-L, Sorrells ME, Raman B, Cairns JE, Tarekegne A, Semagn K, et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda) **2**: 1427–1436

Wu, D., Tanaka, R., Li, X., Ramstein, G.P., Cu, S., Hamilton, J.P., Buell, C.R., Stangoulis, J., Rocheford, T., and Gore, M.A. (2021). High-resolution genome-wide association study pinpoints metal transporter and chelator genes involved in the genetic control of element levels in maize grain. G3 **11**.

Yan, J., Shah, T., Warburton, M.L., Buckler, E.S., McMullen, M.D., and Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS One **4**: e8451.

Zuiderveen, G.H., Padder, B.A., Kamfwa, K., Song, Q., and Kelly, J.D. (2016). Genome-Wide Association Study of Anthracnose Resistance in Andean Beans (*Phaseolus vulgaris*). PLoS One **11**: e0156391.

CHAPTER 2

GENOME SEQUENCING OF FOUR CULINARY HERBS REVEALS TERPENOID GENES UNDERLYING CHEMODIVERSITY IN THE NEPETOIDEAE

This chapter was published in the following manuscript:

Bornowski N, Hamilton JP, Liao P, Wood JC, Dudareva N, Buell CR (2020) Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae. *DNA Res*. doi: 10.1093/dnares/dsaa016.

Bornowski, N, Hamilton, JP, Liao, P, Wood, JC, Dudareva, N, and Buell, CR (2020a) Corrigendum to: Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae. *DNA Res*. doi: 10.1093/dnares/dsaa025.

ABSTRACT

Species within the mint family, Lamiaceae, are widely used for their culinary, cultural, and medicinal properties due to production of a wide variety of specialized metabolites, especially terpenoids. To further our understanding of genome diversity in the Lamiaceae and to provide a resource for mining biochemical pathways, we generated high-quality genome assemblies of four economically important culinary herbs, namely, sweet basil (*Ocimum basilicum* L.), sweet marjoram (*Origanum majorana* L.), oregano (*Origanum vulgare* L.), and rosemary (*Rosmarinus officinalis* L.), and characterized their terpenoid diversity through metabolite profiling and genomic analyses. A total 25 monoterpenes and 11 sesquiterpenes were identified in leaf tissue from the four species. Genes encoding enzymes responsible for the biosynthesis of precursors for mono- and sesqui-terpene synthases were identified in all four species. Across all four species, a total of 235 terpene synthases were identified, ranging from 27 in *O. majorana* to 137 in the tetraploid *O. basilicum*. This study provides valuable resources for further investigation of the genetic basis of chemodiversity in these important culinary herbs.

INTRODUCTION

The Lamiaceae (mint) family is among the largest angiosperm families, containing approximately 7000 species that occupy a wide geographic distribution (Harley et al., 2004) and are commonly recognized by their square stems, opposite leaves, and lobed inflorescences. The Lamiaceae is not only rich in species number and diversity, but also in the production of specialized metabolites. The largest class of these metabolites, terpenes, exhibit substantial chemodiversity and broad variation in abundances across Lamiaceae (Daferera et al., 2002; Mint Evolutionary Genetics Consortium, 2018). Recent molecular-based phylogenetic analyses support ten to twelve major clades within the mint family, the largest of which is the Nepetoideae that contains approximately half of all Lamiaceae species (Li et al., 2016; Mint Evolutionary Genetics Consortium, 2018). The Nepetoideae also has the greatest diversity of monoterpenes (El-Gazzar and Watson, 1970) among the mint clades, making it a robust clade to study the relationship between terpenoid diversity and species diversity.

All terpenoids are derived from two universal precursors, isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP), which are synthesized in plants via two independent pathways: the methylerythriol phosphate (MEP) pathway in the plastid and the mevalonic acid (MVA) pathway distributed among the cytosol endoplasmic reticulum and peroxisomes. IPP and DMAPP then serve as substrates for short-chain prenyltransferases, which produce the prenyl diphosphates, geranyl diphosphate (GPP) in the plastid and farnesyl diphosphate (FPP) in the cytosol. Finally, GPP and FPP are converted to monoterpenes and sesquiterpenes, respectively, by the action of enzymes of the terpene synthase (TPS)

superfamily. Product promiscuity of TPSs and enzymes modifying TPS products are the main sources of terpene structural diversity in plants (Croteau et al., 2000).

A number of Nepetoideae species are used as culinary herbs due to their production of specialized metabolites that impart unique flavor profiles. Sweet basil (*Ocimum basilicum* L.), for example, exhibits a wide range of phenotypes and chemotypes(Simon et al., 1999) and is commonly used in pesto sauce. Although ploidy varies within *O. basilicum*, the cultivar 'Genovese' is tetraploid (2n=4x=48) with a genome size of 4.1-4.7 Gb estimated by flow cytometry (Carović-Stanko et al., 2010; Rewers and Jedrzejczyk, 2016). Rosemary (*Rosmarinus officinalis* L.), in contrast, is an evergreen shrub with grey-green needle-like leaves that emit a strong fragrance due to multiple aromatic volatiles (Zaouali et al., 2010; Jamshidi et al., 2009). *R. officinalis* and its essential oils have been widely used in cultural practices, culinary flavorings, medicinal remedies, and pest deterrents (Sasikumar, 2012). Oregano (*Origanum vulgare* L.) and sweet marjoram (*Origanum majorana* L.) are both members of the *Origanum* genus, having a shrub-like architecture and small, ovate leaves. Their relatedness is evidenced biologically and culturally as *Origanum* spp. can make interspecific hybridizations (Ietswaart, 1980) and some regions of the world refer to oregano and marjoram interchangeably. Young leaves of both *Origanum* spp. are harvested, dried, ground, and added to culinary dishes to bestow a spicy, bitter-sweet flavor. Three major chemotypes have been described for *O. vulgare*: acyclic, cymyl, and sabinyl (Lukas et al., 2015). Chemotypes of *O. majorana* are less-defined due to nomenclature challenges and the presence of distillation artefacts (Fischer et al., 1987, 1988) but both cymyl and sabinyl compounds have been widely-documented (Komaitis et al., 1992; Skoula et al., 1999).

All four species were part of the 1k Plant Transcriptome Initiative (Carpenter et al., 2019; Leebens-Mack et al., 2019) and leaf transcriptomes were generated to examine the evolution of green plants as well as for understanding the evolution of chemodiversity within the Lamiaceae (Mint Evolutionary Genetics Consortium, 2018). In addition, for *O. basilicum*, transcriptomes were generated for cultivars 'Tiguillo' and 'Red Rubin'(Torre et al., 2016) as well as 'CIM Saumya'(Rastogi et al., 2014), and a genome assembly of the *O. basilicum* cultivar 'Perrie' has recently been reported although the actual sequence is not currently available (Dudai et al., 2018). There is a growing number of Lamiaceae species with assembled genomes (see Supplementary Dataset 2.1) that have enabled identification of genes involved in specialized metabolism and a broader understanding of genome organization with respect to specialized metabolism (Lichman et al., 2020b). However, the genetic repertoire encoding chemical diversity within the culinary herbs remain largely unexplored. Here, we here report the genome sequence, annotation, and metabolite profiling of four culinary herbs and describe their repertoire of terpenoid biosynthetic genes that will provide a resource for data-mining not only terpenoid biosynthetic pathway genes but also other genes that function in specialized metabolism.

MATERIALS AND METHODS

Plant Materials and Growing Conditions

Plant samples for *O. basilicum* 'Genovese' and *R. officinalis* 'Arp' were purchased from VanAtta's Greenhouse and Flower Shop (Haslett, MI), whereas *O. majorana* and *O. vulgare* were obtained from Richter's Herbs (Canada). *O. basilicum* and *R. officinalis* were grown in a growth chamber under a 14/10 hour day/night cycle with a daytime temperature of 27 °C and a nighttime temperature of 15 °C; light intensity in the chamber was 210 μE m^{-2} s^{-1}. *O. majorana*

and *O. vulgare* were grown in a greenhouse under a 15/9 hour day/night cycle at a temperature of 26.6 °C. Plant management included weekly fertilizing and pesticide application as necessary. Flow cytometry was performed on leaf samples at the Benaroya Research Institute (Seattle, WA), and k-mer estimated genome sizes were determined using Jellyfish v2.2.0 (Marçais and Kingsford, 2011) with a k-mer size of 31 and adjusted for heterozygous sequences using the R package findGSE v0.1.0 (Sun et al., 2018) .

DNA and RNA Isolation

Nuclei were isolated from leaf tissue following a previously-described protocol (Workman et al., 2018) with an input of 1-2 grams of ground tissue; spin speeds were used based on estimated genome size, 2,700 g, 3,030 g, 3,030 g, and 2,900 g for *O. basilicum*, *O. majorana*, *O. vulgare*, and *R. officinalis*, respectively. DNA was extracted using the Nanobind Plant Nuclei Big DNA Kit (Circulomics, Baltimore, MD, Cat # NB-900-801-01). RNA was extracted from mature leaf tissue using a hot phenol protocol(Davidson et al., 2011) and DNA was removed using the TURBO DNA-*free*™ Kit (Invitrogen, Carlsbad, CA, Cat # AM1907). Quality and concentrations were verified by Nanodrop, Qubit, and agarose gel electrophoresis.

Library Construction, Sequencing, and Expression Abundance Estimation

Genomic libraries were constructed using 10x Genomics Technology (Chromium™ Genome Library Kit & Gel Bead Kit v2; Pleasanton, CA) and sequenced at the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign. Sequencing was performed on an Illumina NovaSeq 6000 at 150 nt in paired-end mode. Libraries were pooled with an aim of 65x coverage for each species. RNA-Seq libraries were constructed using the

Illumina TruSeq Stranded mRNA Kit with polyA mRNA selection and IDT for Illumina Unique Dual Index (UDI) primers (Illumina, San Diego, CA) and sequenced at the Michigan State University Research Technology Support Facility. RNA-Seq libraries were sequenced on an Illumina HiSeq 4000 at 150 nt in paired-end mode. Reads were cleaned with Cutadapt v2.3 (Martin, 2011) which trimmed adapters and 3' bases with a quality score less than 10, and only kept reads of at least 100 nt. After cleaning, reads were aligned to their respective genomes using HISAT2 v2.1.0 (Kim et al., 2019) with the following parameters set: --dta-cufflinks, --max-intronlen 5000, and --rna-strandness RF. Cufflinks v2.2.1(Trapnell et al., 2010) was run in stranded mode to generate expression abundances (fragments per kb exon model per million mapped reads, FPKM).

Genome Assembly and Annotation

10x Genomics reads were demultiplexed and assembled using Supernova v2.1.1 (Weisenfeld et al., 2017) with --maxreads set to 900 million, 330 million, 259 million, and 450 million reads for *O. basilicum*, *O. majorana*, *O. vulgare*, and *R. officinalis*, respectively. Scaffolds containing only N sequences were removed from the final assemblies. Custom repeat libraries (CRL) were generated for each species using RepeatModeler v1.0.8 (http://www.repeatmasker.org) as described previously (Zhao et al., 2019). Genome assemblies were masked with their respective CRLs using RepeatMasker v4.0.6 (http://www.repeatmasker.org). Gene prediction on the masked assembly was performed using Augustus v3.1 (Stanke et al., 2008) with a matrix trained for the Nepetoideae species, *Hyssopus officinalis* L. (Lichman et al., 2020a). To refine the gene models, leaf RNA-Seq libraries were cleaned and used to generated genome-guided transcript assemblies using Trinity v2.3.2

(Grabherr et al., 2011) with a maximum intron size of 5000 and a minimum contig length of 500 nt in stranded mode. The genome-guided transcript assemblies were used with PASA2 v2.3.3 (Campbell et al., 2006) to create the working gene model set. To identify high confidence gene models from the working gene model set, the gene models were searched against PFAM v32.0 (El-Gebali et al., 2018) using HMMER v3.2.1 (hmmer.org) with search cutoffs --domE 1e-3 -E 1e-5, and gene abundances of the leaf RNA-Seq library were calculated using Kallisto v0.46.0 (Bray et al., 2016). Gene models that were not partial models, did not contain an internal stop codon, not transposable element-related, and had a PFAM domain match or a TPM > 0 were selected as high confidence models. Functional annotation of the high confidence gene models was generated as described previously (Zhao et al., 2019). Gene ontology terms were assigned to the representative high confidence gene models using IPRscan v5.34.73.0 (Jones et al., 2014).

Genome Sequence and Annotation Quality Assessment

To assess the completeness of the assembly, whole genome shotgun libraries were processed using Cutadapt v2.3 (Martin, 2011) and reads were aligned to their respective assemblies with BWA-MEM v0.7.16a (Li and Durbin, 2009). Paired-end RNA-Seq libraries constructed from leaf tissue were aligned to the assemblies using HISAT2 v2.1.0 (Kim et al., 2019) using stranded mode with a maximum intron length of 5000 bp. Coverage of the genic space was assessed using BUSCO v3.0.2b (Waterhouse et al., 2018; Simão et al., 2015) with the Embryophyta odb9 dataset (creation date: 2016-02-13, number of species: 30, number of BUSCOs: 1440) to detect conserved orthologs in the assemblies.

Extraction and Analysis of Terpenoids by GC-MS

The same leaf tissue harvested for RNA extraction was used for terpenoid profiling. Tissue from each species (0.2 g) was ground in liquid nitrogen and extracted overnight with shaking at room temperature with 5 mL of dichloromethane containing 6.6 µg of the internal standard naphthalene. After centrifugation, the solvent containing the extracted metabolites was transferred to a new glass tube and concentrated to ~180 µl under nitrogen gas (Dudareva et al., 2005). Subsequently, GC-MS analysis was performed on an Agilent 6890 gas chromatograph (Agilent Technologies) equipped with a HP-5MS column (30m, 0.25 mm, 0.25 µm; Agilent Technologies) and coupled to an Agilent 5975B insert MSD quadrupole mass spectrometer (Agilent Technologies). Each sample (2 µL) was injected at a pulsed splitless mode at 250°C. The column temperature was held at 50°C for 2 min, followed by increased to 320°C at 20°C min-1, and held at 320°C for 4.5 min. Helium was applied as a carrier gas at a flow rate of 1 mL min-1. MS ionization energy was set at 70 eV, and the mass spectrum was scanned from 50 to 300 amu. Three biological replicates were used for metabolite profile analysis for each species. Compounds were identified by comparing retention times and mass spectra with those of commercially available authentic standards including α-pinene, β-pinene, α-phellandrene, α-terpinene, *cis*-β-ocimene, γ-terpinene, terpinolene, linalool, geraniol, β-caryophyllene, and caryophyllene oxide as well as by comparing mass spectra to the National Institute of Standards and Technology (NIST) Mass Spectral Library v2.2. Quantification of terpenoids was performed using the Mass Hunter quantitative software (Agilent Technologies, v.B. 07.01) using response factors relative to the internal standard determined experimentally for the commercially available authentic standards α-pinene (representative monoterpene for α-thujene, α-pinene, camphene), β-pinene (representative monoterpene for β-pinene and β-myrcene), α-phellandrene (representative

monoterpene for α-phellandrene and β-phellandrene), α-terpinene (representative monoterpene for α-terpinene, γ-terpinene, o-cymene, *cis*-β-ocimene, terpinolene), geraniol (representative monoterpene alcohol), β-caryophyllene (representative sesquiterpene), and nerolidol (representative sesquiterpene alcohol) and normalized to the fresh weight of the tissue.

Comparative Genome Analyses

Representative peptide transcripts from teak (*Tectona grandis* L.f.) (Zhao et al., 2019) and *Arabidopsis thaliana* Col-0 (Cheng et al., 2017) were included in comparative genome analysis as outgroups for Nepetoideae and Lamiaceae, respectively. Predicted teak peptides were downloaded from GigaDB (http://dx.doi.org/10.5524/100550) on 26Nov19 and *A. thaliana* peptides were downloaded from Araport11 on 13Nov19. Orthofinder2 v2.3.7 (Emms and Kelly, 2018a) was run using default settings to identify orthologous and paralogous TPSs in each species. Orthologous groups represented by all species were used to construct and root a consensus species tree with the STAG (Emms and Kelly, 2018b) and STRIDE (Emms and Kelly, 2017) algorithms, respectively. Gene family expansion and contraction was determined with CAFE v4.2.1 (Han et al., 2013) using default settings and an ultrametric tree rooted at 125 million years ago based on estimates from multiple studies (Zeng et al., 2017; Magallón et al., 2015; Bell et al., 2010; Dong et al., 2018). Enrichment of gene ontology terms (GO terms) was performed using the Bioconductor package TopGO v2.38.1 (Alexa and Rahnenfuhrer, 2019).

Identification of TPS Orthologs

Manually-reviewed cloned terpene synthase (TPS) genes from Lamiaceae were retrieved from SwissProt (Supplementary Table S2.1). TPSs were selected to include species within and

outside the Nepetoideae subfamily of interest, as well as discrete clades within the Nepetoideae. Lamiaceae TPSs were used along with annotated *A. thaliana* TPSs to identify putative orthologs in the predicted proteomes of the four culinary herbs.

Data Availability

Raw sequences are available in the National Center for Biotechnology Information Sequence Read Archive under BioProject PRJNA592145. Large files associated with the genomes including genome sequence, annotation, gff and expression matrices are available in the Dryad Digital Repository under doi (https://doi.org/10.5061/dryad.jwstqjq6t).

RESULTS AND DISCUSSION

Genome Assembly and Annotation

Flow cytometry of the four culinary herbs revealed estimated haploid genome sizes consistent with previous studies. The flow cytometry haploid genome estimation of the tetraploid *O. basilicum* var. 'Genovese' was 2.34 Gb which is within previous estimates of 2.04 Gb (Carović-Stanko et al., 2010) to 2.37 Gb (Rewers and Jedrzejczyk, 2016). The flow cytometry estimate of haploid genome size for *O. majorana* (880.2 Mb) is comparable to a recent estimate of 846 Mb (Jedrzejczyk, 2018), and estimates for *O. vulgare* (694.38 Mb) and *R. officinalis* (1198.05 Mb) are remarkably similar to previously-published flow cytometrical estimates of 684.6 Mb (Mowforth, 1985) and 1198.05 Mb (Pellicer et al., 2010), respectively. The k-mer estimated genome sizes for *O. basilicum*, *O. majorana*, *O. vulgare*, and *R. officinalis* were 2.15 Gb, 760.95 Mb, 665.08 Mb, and 1013.85 Mb, respectively (Supplementary Table S2.2; Supplementary Figure S2.1); overall, estimation of genome sizes between flow cytometry and k-

mer frequency were comparable. We utilized Supernova (Weisenfeld et al., 2017) to assemble the genomes of the four species. As shown in Table 2.1, reconstituted molecule lengths ranged from 36.91 kb (*O. vulgare*) to 83.47 kb (*R. officinalis*). Assembled contig N50 lengths ranged from 21.82 kb (*R. officinalis*) to 48.30 kb (*O. basilicum*), while scaffold N50s ranged from 368.74 kb (*R. officinalis*) to 1.51 Mb (*O. basilicum*). The GC content of the assemblies varied from 38.12% (*R. officinalis*) to 40.32% (*O. vulgare*). Detection of heterozygous SNPs by the Supernova assembly process ranged from 192 bp to 1490 bp in *R. officinalis* and *O. basilicum*, respectively. For the final genome assembly, scaffolds less than 10 kb were removed resulting in final assembly sizes of *O. basilicum* (2.07 Gb), *R. officinalis* (1.01 Gb), *O. majorana* (760.89 Mb), and *O. vulgare* (630.04 Mb).

Although the Supernova assembler was originally designed for human genomics applications, it has been used to assemble non-human animal species like perch (Ozerov et al., 2018) and rice coral (Helmkampf et al., 2019), as well as diploid plant species such as pepper (Hulse-Kemp et al., 2018), snowberry (Lau et al., 2020), and maize (Ott et al., 2018). Supernova assemblies have been generated for polyploid species including proso millet (*Panicum miliaceum*), an allotetraploid (Ott et al., 2018), and potato (*Solanum tuberosum subsp. andigena*), an autopolyploid (Kyriakidou et al., 2020). Our successful assembly of *O. basilicum* further supports the use of Supernova to generate quality assemblies of polyploid species.

To assess the completeness and representation of genic sequences, whole genome shotgun and RNAseq reads were aligned to their cognate genome assembly. At least 95.8% of whole genome shotgun reads aligned to the assemblies (Supplementary Table S2.3); properly

paired reads with correct orientation ranged from 79.0% (*R. officinalis*) to 85.1% (*O. basilicum*). Leaf RNA-Seq reads had overall alignment rates ranging from 82.7% (*O. basilicum*) to 89.4% (*O. majorana*) (Supplementary Table S2.4). BUSCO analysis of the *O. majorana*, *O. vulgare*, and *R. officinalis* assemblies revealed 89.5% to 90.1% complete orthologs while the *O. basilicum* assembly contained 86.7% of complete orthologs (Table 2.2). Approximately half of the *O. basilicum* orthologs were present as multiple copies, consistent with its tetraploidy (Carović-Stanko et al., 2010).

Of the species assembled in this study, a previous assembly was reported for *O. basilicum* cultivar, 'Perrie' (Dudai et al., 2018) while the present study assembled the cultivar 'Genovese.' Both cultivars are tetraploid (2n=4x=48) yet flow cytometry estimates of haploid genome size differ; 'Perrie' was estimated at 1.59 Gb (Koroch et al., 2010), while our flow cytometry estimate of 'Genovese' was 2.34 Gb is in agreement with previous estimations (Carović-Stanko et al., 2010; Rewers and Jedrzejczyk, 2016) and estimated size using k-mer frequency as well as from the Supernova program. The 'Perrie' assembly was 2.13 Gb and the 'Genovese' assembly size generated in this study was 2.07 Gb; these differences may reflect variation in genome size among and within *Ocimum spp.* (2010) as well as the level of heterozygosity in the sequenced genomes. Dudai et al. (2018) report 'Perrie' as highly homozygous while our 'Genovese' sample was heterozygous. The 'Genovese' N50 contig length was slightly larger than the Dudai et al. 'Perrie' assembly (48.30 kb to 45.71 kb, respectively), although the N50 scaffold size was substantially smaller (1.51 Mb to 19.30 Mb) (Supplementary Table S2.5). Assessment of genic completeness using 1440 BUSCO genes revealed 93.0% and 86.7% of complete genes in

'Perrie' and 'Genovese', respectively; however, our 'Genovese' assembly contained 30.5% of these genes as single-copy compared to 18.5% for the 'Perrie' assembly.

The four genomes were annotated using the gene finder Augustus (Stanke et al., 2008) and the resulting gene models were refined with PASA2 (Campbell et al., 2006) using the genome-guided transcript assemblies. The initial working gene model sets were filtered for high confidence genes using expression evidence and/or PFAM domains, resulting in high confidence gene models for *O. basilicum* (n=78,990), *O. majorana* (n=33,929), *O. vulgare* (n=32,623), and *R. officinalis* (n=51,389) (Supplementary Table S2.6). The annotation data sets were assessed for completeness using BUSCO, revealing a high proportion of complete single copy orthologs. Specifically, single copy orthologs were considered complete in the high confidence representative gene model set at frequencies of 88.3% (*O. basilicum*), 90.6% (*O. majorana*), 89.7% (*O. vulgare*), and 89.2% (*R. officinalis*). As expected, the tetraploid *O. basilicum* gene model sets contained substantially more duplicated orthologs compared to the three other diploid species.

Repeat-masking was performed on the culinary herb genome assemblies to mask repetitive elements (Supplementary Table S2.7). The proportion of masked bases was not dependent on genome assembly size, as *R. officinalis* (54.7%) and *O. basilicum* (61.6%) had a similar proportion of masked bases compared to *O. majorana* (65.5%) and *O. vulgare* (65.4%). Long terminal repeats (LTRs) were the most common repetitive elements, though the proportion of LTRs varied among the culinary herbs. LTRs represented 31.6% of the *R. officinalis* assembly compared to 47.8% and 49.3% of the *O. majorana* and *O. vulgare* assemblies. DNA elements

were the second-most common classified element and represented 3.9% of the *O. vulgare* assembly up to 5.3% of the *R. officinalis* assembly. The proportion of long interspersed nuclear elements (LINEs) identified in the assemblies ranged from 0.28% (*O. majorana*) to 1.2% (*O. vulgare).* Instead, *O. majorana* had noticeably higher amounts of short interspersed nuclear elements (SINEs) identified (n=3,469) compared to the other assemblies containing 209, 420, and 600 SINEs for *R. officinalis*, *O. basilicum*, and *O. vulgare*, respectively. The number of satellite repeats was associated with assembly size, though they did not represent greater than 1.1% of the assembly size in any culinary herb.

Mono- and Sesqui-terpene Profiles of Culinary Herbs

As terpenoids are produced primarily in the leaves (Turner and Croteau, 2004), metabolite profiling of leaf terpenoids from the four species was performed by gas chromatography-mass spectrometry (GC-MS) (Figure 2.1). Spectrometric analyses revealed that these plants produce both monoterpenes and sesquiterpenes, with monoterpenes contributing to a higher degree (Supplementary Table S2.8). A total of 25 different monoterpenes were identified with only β-myrcene produced in all cultivars. Ten monoterpenes were species-specific (for example, carvacrol was found only in *O. vulgare*), while the others, such as γ-terpinene, were shared by two or three species. The highest monoterpene diversity was detected in *R. officinalis*, which produced 17 monoterpenes, while the other cultivars synthesized 10-11 compounds. The total amounts of produced monoterpenes also varied between the species, ranging from 8.99 μmol g FW^{-1} in *O. vulgare* to just 0.52 μmol g FW^{-1} in *O. majorana* (Supplementary Table S2.8). The obtained metabolic profiles were generally consistent with literature reports (Sivropoulou et al., 1996; Daferera et al., 2000; Iijima et al., 2004; Crocoll et al., 2010).

In contrast to rich chemical diversity observed for monoterpenes, the amount and spectrum of sesquiterpenes was significantly lower. A total of eleven sesquiterpenes were detected, five of which were unique to a single species. There was no sesquiterpene shared by all four species, although β-caryophyllene was produced by three species. While *O. basilicum* produced the most diverse spectrum of sesquiterpenes, the highest amount of sesquiterpenes was found in *O. vulgare*, suggesting that this species is the highest producer of both mono- and sesquiterpenes. Comparative analysis of the most abundant compounds revealed that in species with relatively high levels of terpenoids, *O. vulgare* and *R. officinalis*, these are mostly monoterpenes, while in low terpene producers, such as *O. basilicum* and *O. majorana*, sesquiterpenes contribute to the overall terpenoid profile. This analysis also revealed that the spectra of most abundant compounds are mostly species-specific (Figure 2.2).

Orthologous and Paralogous Clustering

Orthofinder2 is a software program that partitions genes according to their phylogenetic ancestry (Emms and Kelly, 2018a) and clusters them into orthologous (orthogroups) and paralogous clusters. Identification and comparison of orthologs within orthogroups may reveal gene duplication or loss over evolutionary time. Thus, we performed this type of analysis for the four Nepetoideae species used in this study along with two additional species included in the analysis as outgroups. *T. grandis* (teak) was used as a non-Nepetoideae Lamiaceae species along with the model species *A. thaliana*, a member of the Brassicaceae. High confidence representative predicted peptides for these six species, along with curated Lamiaceae terpene synthases obtained from SwissProt, were used as input for Orthofinder2; in total, 219,047

predicted peptides were included. Of these, 200,920 (91.7%) were assigned to 25,660 orthogroups (Figure 2.3A; Supplementary Dataset 2.2).

Orthogroup occupancy by species was similar for the Lamiaceae species, ranging from 65.5% (*T. grandis*) to 76.6% (*O. basilicum*), while orthologous genes from the non-Lamiaceae outgroup *A. thaliana* were only present in 53.9% of orthogroups. A rooted species tree (Emms and Kelly, 2017, 2018b) revealed a topology in agreement with a previous Lamiaceae cladogram (Figure 2.3B) (Mint Evolutionary Genetics Consortium, 2018). As expected, *O. majorana* and *O. vulgare* were closely related, and teak and *A. thaliana* were more distantly related to the rest of the Nepetoideae species.

In total, 10,407 orthologous groups contained orthologs from all six species (Supplementary Dataset 2.2). Genes in these groups were enriched in core biological processes and molecular function such as translation (GO:0006412; p<1e-30), intracellular protein transport (GO:0006886; p<1e-30), and structural constituent of ribosome (GO:0003735; p<1e-30). Lamiaceae members shared 2,368 orthologous groups containing genes involved in oxidation-reduction (GO:0055114; p<1e-30), protein phosphorylation (GO:0006468; p<1e-30), regulation of transcription (GO:0006355; p<1e-30), defense response (GO:0006952; p<1e-30), and terpene synthase activity (GO:0010333; p=2.3e-13), among others. Of the orthologous genes unique to the culinary herbs and their singletons (Supplementary Table S2.9), the most-significant biological processes were recognition of pollen (GO:0048544; p=1.5e-28) and translation (GO:0006412; p=3.6e-06). These genes were associated with cellular locations such as the ribosome (GO:0005840; p=1.7e-07) and nucleosome (GO:0000786; p=0.00056). The most

significant molecular functions for this subset of genes were protein serine/threonine kinase activity (GO:0004674; p=7.3e-16), ADP binding (GO:0043531; p=1.2e-14), and terpene synthase activity (GO:0010333; p=5.80e-11). Considering that divalent metal cofactors have been shown to influence terpene synthase activity and specificity (Köllner et al., 2004, 2008), other notably enriched terms among the culinary herb-specific orthologous genes included magnesium ion binding (GO:0000287; p=5.6e-05), manganese ion binding (GO:0030145; p=0.00068), and copper ion binding (GO:0005507; p=0.00994).

Gene Family Analysis

Of the 25,660 orthologous groups identified by the Orthofinder analysis, 12,615 were inferred to be present in the most recent common ancestor and were used in the CAFE analysis along with the ultrametric tree. The number of gene families in the observed Lamiaceae species was found to be generally consistent over evolutionary time (Figure 2.3C). Compared to the rest of the culinary herbs, a noticeable gene family contraction occurred in the *Origanum* genus, while *O. basilicum* shows significant gene family expansion, likely due in part to its tetraploidy. Both non-Nepetoideae outgroups share a similar number and proportion of contracted gene families.

To better understand evolutionary relationships of these four culinary herbs among the ever-growing list of sequenced Lamiaceae spp., we conducted an Orthofinder analysis for all available Lamiaceae predicted proteomes (Supplementary Dataset 2.1; Supplementary Figure S2.2), characterizing 34,998 orthologous groups. Approximately 30% of the orthologous groups contained at least one ortholog from each species. Intra-genus orthologous groups for the

Origanum spp. and *Nepeta* spp. contained the second- and third-most number of non-encompassing intersections. Orthologs from *Pogostemon cablin*, an octoploid, were represented in the most orthologous groups. The species tree derived from ancestral gene families was in agreement with previous cladograms (Mint Evolutionary Genetics Consortium, 2018; Li et al., 2016, 2017), confirming the monophyly of Nepetoideae subfamily and the polyphyly of the *Salvia* genus, as described previously (Walker et al., 2004).

Identification of Precursor Genes and Terpene Synthases in Four Culinary Herbs

Terpenes are synthesized from common IPP and DMAPP precursors via the MEP and MVA pathways. To examine the terpenoid biosynthetic pathway in the culinary herbs, orthologous groups were queried for genes belonging to *A. thaliana* MEP and MVA pathways, as identified previously (Supplementary Table S2.10) (Mint Evolutionary Genetics Consortium, 2018). All 22 of these *A. thaliana* MEP/MVA genes clustered into 17 orthologous groups (Supplementary Table S2.11). Six additional *A. thaliana* genes also clustered with the MEP/MVA orthogroups OG0001733 and OG0006021, representing five geranylgeranyl phosphate synthases and a putative 1-deoxyxylulose 5-phosphate synthase, respectively. A total of 148 culinary herb orthologs were present among the MEP/MVA orthologous groups. In 13 of the 17 orthogroups, each culinary herb contained equal to or greater numbers of orthologs than *A. thaliana*. However, the difference in the number of MEP/MVA orthologs across all species was small as each culinary herb contained one to ten MEP/MVA orthologs, compared to one to six orthologs in *A. thaliana* and one to three orthologs in *T. grandis* (Figure 2.4A).

Terpene synthase enzymes synthesize terpenoids from the GPP and FPP end products of the MEP and MVA pathways. To identify TPS genes in the four culinary herbs, orthogroup occupancy of previously published *A. thaliana* TPSs (Mint Evolutionary Genetics Consortium, 2018) as well as curated Lamiaceae TPSs was investigated. The *A. thaliana* TPSs (n=34) belong to TPS subfamilies TPSa, TPSb, TPSc, TPSe/f, and TPSg; these genes clustered into twelve orthologous groups and six singletons (Table 2.3; Supplementary Table S2.12). Six of the twelve orthologous groups containing *A. thaliana* TPSs were unique to *A. thaliana*; the other six orthologous groups contained a total of 151 putative TPSs from the four culinary herbs. Curated Lamiaceae TPSs (n=26) were also included in the analyses to identify Lamiaceae-specific TPSs (Supplementary Table S2.1); these "bait" Lamiaceae TPS genes clustered into seven orthologous groups containing a total of 212 putative terpene synthases across the culinary herbs (Table 2.3). Within these orthogroups, *O. basilicum* contained the most TPSs (n=128) of the culinary herbs, followed by *R. officinalis* (n=35), *O. vulgare* (n=26), and *O. majorana* (n=23). In these same orthogroups, 57 orthologs from *T. grandis* were detected along with eight orthologs from *A. thaliana*. Orthologous groups OG0000008, OG0000079, and OG0001915 contained TPSs from both *A. thaliana* and Lamiaceae bait TPSs, representing TPS subfamilies TPSa, TPSb, and TPSc, respectively.

Overall, the culinary herb genomes encoded a total of 235 putative terpene synthases occupying ten orthologous groups with terpene synthases from *A. thaliana* or the Lamiaceae bait. Nearly four to five times as many TPSs were found in *O. basilicum* (n=137) compared to the diploid culinary herbs that contained substantially fewer TPSs: *R. officinalis* (n=38), *O. vulgare* (n=33), *O. majorana* (n=27) (Figure 2.4B). Consistent with the relatively high levels and rich

chemical diversity of monoterpenes in these culinary herbs (Supplementary Table S2.8), these TPSs were mainly represented by members of the TPSb subfamily, which includes most of the angiosperm monoterpene synthases (Chen et al., 2011).

Among the six orthogroups jointly occupied by *A. thaliana* TPSs and culinary herb orthologs, five orthogroups contained Lamiaceae orthologs in the same or greater quantity than the *A. thaliana* TPSs (Table 2.3). Whereas *A. thaliana* contained one to six orthologs in both the MVA/MEP and TPS related orthogroups, the culinary herbs generally contained more TPS orthologs compared to MEP/MVA orthologs. (Figure 2.4; Supplementary Table S2.11). For example, orthogroup OG0000008 contained six *A. thaliana* TPSs, all belonging to the TPSb subfamily, while all other mint species were represented by twelve (*O. majorana* and *R. officinalis*) to 38 (*O. basilicum*) orthologs. In OG0000079, a single *A. thaliana* sesquiterpene synthase gene, TPS21 (AT5G23960) was present along with one and two orthologs in *O. majorana* and *O. vulgare*, respectively, and 32 orthologs in *O. basilicum*. However, this trend did not hold for all orthologous groups. For example, the highest number of orthologs in orthogroups OG0003155 (n=7) and OG0001915 (n=4) belonged to *T. grandis*; the *A. thaliana* TPSs in these orthogroups were associated with TPSe/f and TPSc subfamilies, respectively. Lineage-specific *A. thaliana* TPS orthogroups included OG0016169 (n=5), OG0018766 (n=3), OG0018926 (n=3), OG0020906 (n=2), OG0020971 (n=2), and OG0021104 (n=2) in addition to the six singletons.

Physical Clustering of Specialized Metabolite Pathways

Physical clustering within the genome has been reported for numerous specialized metabolism biosynthetic pathways (Nützmann et al., 2016), thus to identify putative clusters of

enzymes involved in secondary metabolite synthesis, plantiSMASH v1.0 analysis (Kautsar et al., 2018) was performed on each culinary herb genome assembly with its high confidence representative gene set (Supplementary Table S2.13). The quantity and classification of clusters detected varied by species. In total, there were 104 clusters detected in *O. basilicum*, 36 clusters detected in *O. majorana*, 38 clusters detected in *O. vulgare*, and 22 clusters detected in *R. officinalis*. In particular, the four species were enriched in clusters related to terpene and saccharide production. Among the culinary herbs, the most terpene clusters were found for the *O. basilicum* assembly, with 226 genes located across 24 clusters. The *O. majorana* and *O. vulgare* assemblies contained eight and nine terpene clusters, with 65 and 83 corresponding genes, respectively. The smallest number of terpene clusters was found in *R. officinalis*, with 26 genes in two clusters. In comparison, *A. thaliana* had seven putative terpene clusters containing 129 genes, and *T. grandis* had six clusters with 72 genes. Other secondary metabolite clusters were represented by a combination of terpene-related genes along with other secondary metabolites like alkaloids, lignans, and polyketides.

CONCLUSION

Plants in the mint family are used worldwide for their unique chemical profiles conferred by specialized metabolites such as terpenoids. In this study, we focused on four mint species commonly utilized as culinary herbs: *O. basilicum*, *O. majorana*, *O. vulgare*, and *R. officinalis*. Genome sequencing, assembly, and annotation of these herbs revealed a diversity of genes involved in terpenoid biosynthesis. In addition, targeted metabolic profiling revealed the diversity of monoterpenes and sesquiterpenes in these species and exemplified unique terpenoid profiles for each species. Our study showcases the genomic and metabolomic characterization of these four herbs that can be used to further explore terpene biosynthesis.

ACKNOWLEDGEMENTS

Funding for work on specialized metabolism was provided in part by a grant to C.R.B. and N.D. from the National Science Foundation (IOS-1444499), the USDA Hatch Act (C.R.B), and the USDA National Institute of Food and Agriculture Hatch Project number 177845 (N.D.). We thank Brieanne Vaillancourt with her assistance in sequence data management.

APPENDIX

FIGURES

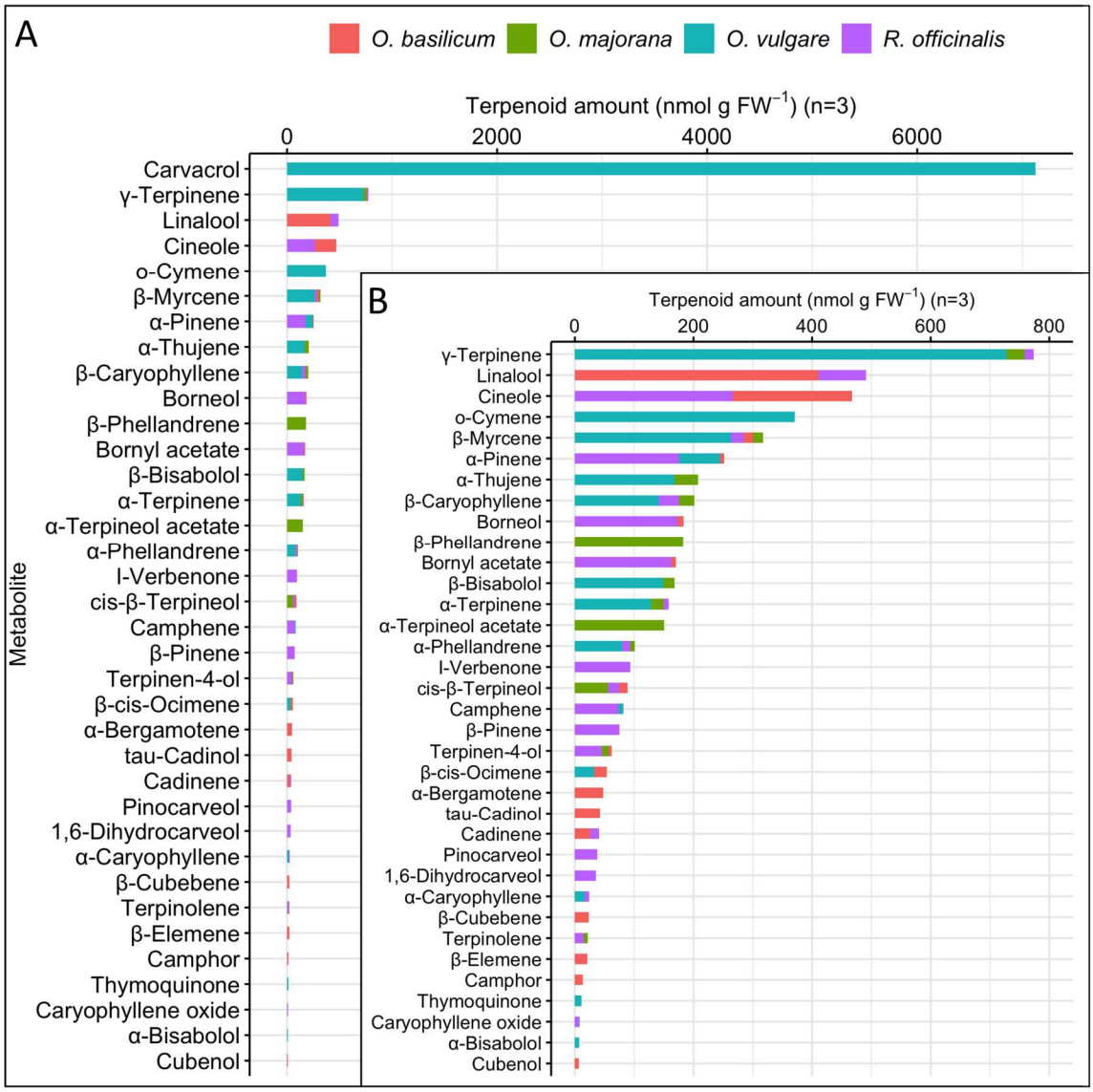


Figure 2.1. Terpenoid profiles in four culinary herbs.
Leaf tissue from four culinary herbs was subjected to targeted metabolite profiling. Terpenoid levels are the average of three replicates, measured in nmol per gram of fresh weight.
A: Distribution of metabolites including carvacrol.
B: Distribution of metabolites excluding carvacrol.

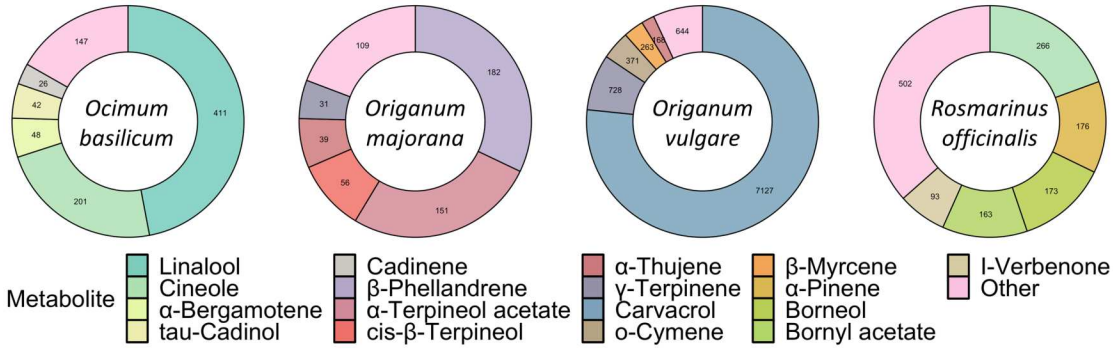


Figure 2.2. Unique terpenoid profiles in leaf tissue of four culinary herbs.
The five most detected terpenoids are indicated for each culinary herb, and remaining terpenoids are classified as "Other." Terpenoid levels are the average of three replicates, measured in nmol per gram of fresh weight.

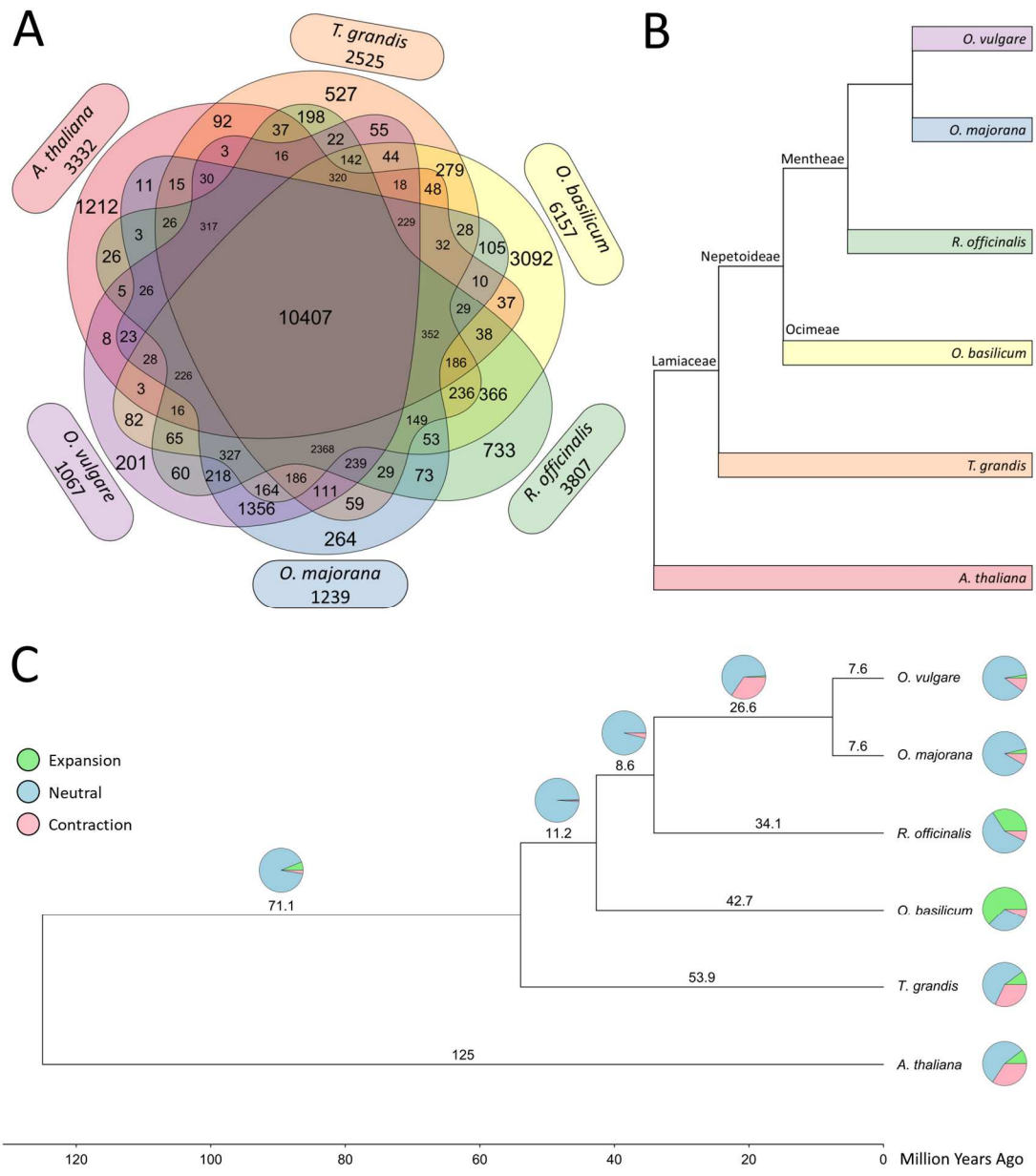


Figure 2.3. Orthologous relationships between four culinary herbs.
A: Venn diagram of orthologous groups shared among species. Numbers next to the species and outside of the plot indicate the number of singleton genes. B: Cladogram of the species' evolutionary relationships. Species belonging to the Lamiaceae family are further divided into the Nepetoideae subfamily and Mentheae and Ocimeae clades (Eltsholtzieae clade is not represented here). C: Gene family evolution of four Nepetoideae culinary herbs and outgroups. Using a root age of 125 million years ago, the species were estimated to share 12,615 gene families. Changes to gene family sizes for each lineage are indicated by pie charts, where green indicates expansion, blue indicates neutrality, and red indicates contraction.

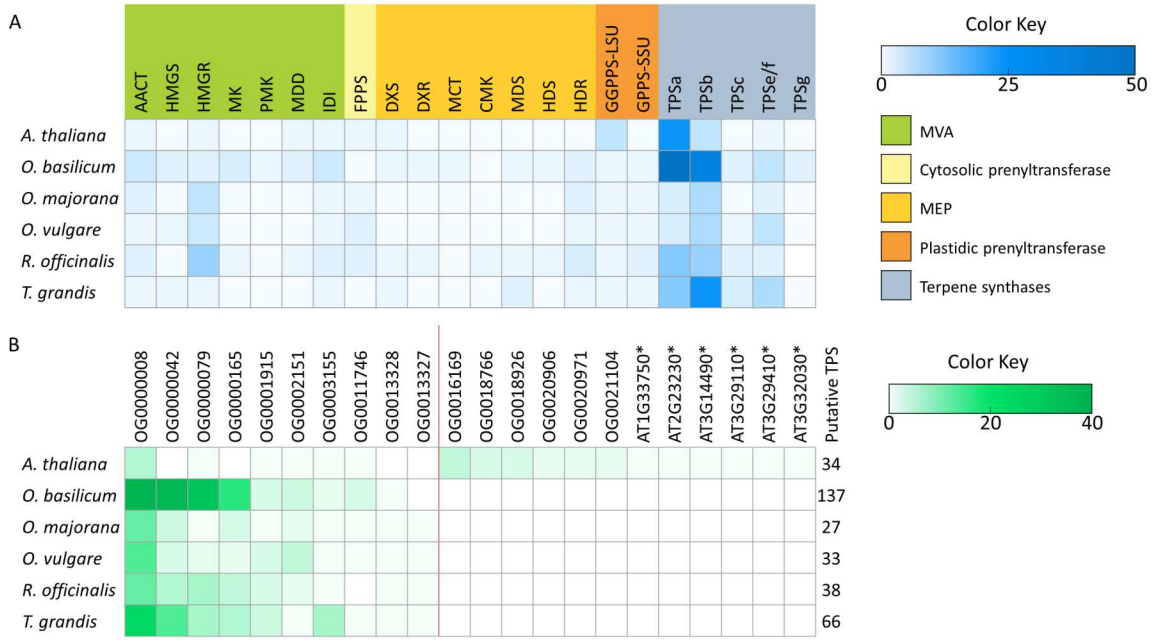


Figure 2.4. Orthologous gene clusters in the terpenoid biosynthetic pathway.
The number of corresponding orthologs from each species is indicated by color gradient intensity.
A: Orthogroups occupied by *Arabidopsis thaliana* genes involved in the mevalonic acid (MVA) and methylerythriol phosphate (MEP) pathways. MVA pathway leading to sterols and sesquiterpenes: acetyl-CoA acetyltransferase activity (AACT); 3-hydroxy-3-methylglutaryl coenzyme A synthase (HMGS); hydroxymethylglutaryl-CoA reductase (HMGR); ATP:mevalonate phosphotransferase (MK); phosphomevalonate kinase (PMK); mevalonate diphosphate decarboxylase (MDD); isopentenyl diphosphate delta-isomerase (IDI); geranyltranstransferase (FPPS).
MEP pathway leading to geraniol, monoterpenes, and diterpenes: 1-deoxy-D-xylulose-5-phosphate synthase (DXS); 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR); 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (MCT); 4-(cytidine 50-diphospho)-2-C-methyl-D-erythritol kinase (CMK); 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS); 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (HDS); 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate reductase (HDR); geranylgeranyl pyrophosphate synthase large subunit (GGPPS-LSU); geranyl pyrophosphate synthase small subunit (GPPS-SSU).
Terpene synthase subfamilies: TPS a, b, c, g, e/f.
B: Orthogroups occupied by *A. thaliana* and Lamiaceae terpene synthase genes. Orthologous groups specific to *A. thaliana* are shown to the right of the red divider line; singletons are denoted by an "*" following the gene ID.

TABLES

Table 2.1. Assembly metrics of four culinary herbs.

	Ocimum basilicum	***Origanum majorana***	***Origanum vulgare***	***Rosmarinus officinalis***
Ploidy	2n = 4x = 48[a]	2n = 2x = 30[b]	2n = 2x = 28,30,32[b]	2n = 20,24[b]
Estimated haploid genome size (flow cytometry) (Mb)	2337.42	880.20	694.38	1198.05
Estimated genome size (Supernova) (Mb)	2360	858.29	705.35	1180
Assembled genome size (Supernova) (Mb)	2067.62	760.89	630.04	1013.85
Mean molecule length (kb)	51.26	43.28	36.91	83.47
Mean distance between heterozygous SNPs (bp)	1490	1340	273	192
Number of scaffolds >= 10 kb	17105	8763	13832	23035
N50 contig size (kb)	48.30	35.95	26.28	21.82
N50 scaffold size (kb)	1506.96	1383.40	157.94	368.74
Assembly GC content	38.42%	40.17%	40.32%	38.12%
Assembly N content	14.65%	11.70%	8.50%	18.67%

[a] Carović-Stanko et al. 2010

[b] Rice et al. 2015

Table 2.2. Representation of genic space in four culinary herb genome assemblies as revealed through Benchmarking Single Copy Orthologs (BUSCO)[a].

Species	Complete	Single copy	Duplicated	Fragmented	Missing	Total
Ocimum basilicum	1248 (86.7%)	439 (30.5%)	809 (56.2%)	39 (2.7%)	153 (10.6%)	1440
Origanum majorana	1289 (89.5%)	877 (60.9%)	412 (28.6%)	30 (2.1%)	121 (8.4%)	1440
Origanum vulgare	1297 (90.1%)	1218 (84.6%)	79 (5.5%)	38 (2.6%)	105 (7.3%)	1440
Rosmarinus officinalis	1297 (90.1%)	1216 (84.4%)	81 (5.6%)	38 (2.6%)	105 (7.3%)	1440

[a] Simão et al. 2015; Waterhouse et al. 2018

Table 2.3. Orthogroup occupancy of *Arabidopsis thaliana* and Lamiaceae terpene synthase genes.

Orthogroup	*A. thaliana*	TPS bait	*O. basilicum*	*O. majorana*	*R. officinalis*	*T. grandis*	*O. vulgare*	Total	Subfamily
OG0000008	6	12	38	12	12	24	14	118	TPSb
OG0000079	1	1	32	1	7	7	2	51	TPSa
OG0001915	1	1	3	1	3	4	3	16	TPSc
OG0000042	0	3	37	4	6	14	3	67	-
OG0000165	0	4	17	3	5	6	2	37	-
OG0013327	0	3	0	1	1	1	1	7	-
OG0013328	0	2	1	1	1	1	1	7	-
OG0002151	1	0	4	2	2	1	5	15	TPSe/f
OG0003155	1	0	2	1	1	7	1	13	TPSe/f
OG0011746	1	0	3	1	0	1	1	7	TPSg
OG0016169	5	0	0	0	0	0	0	5	TPSa
OG0018766	3	0	0	0	0	0	0	3	TPSa
OG0018926	3	0	0	0	0	0	0	3	TPSa
OG0020906	2	0	0	0	0	0	0	2	TPSa
OG0020971	2	0	0	0	0	0	0	2	TPSa
OG0021104	2	0	0	0	0	0	0	2	TPSa
OG0026006	1	0	0	0	0	0	0	1	TPSa
OG0026746	1	0	0	0	0	0	0	1	TPSa
OG0027218	1	0	0	0	0	0	0	1	TPSa
OG0027424	1	0	0	0	0	0	0	1	TPSa
OG0027428	1	0	0	0	0	0	0	1	TPSa
OG0027452	1	0	0	0	0	0	0	1	TPSa

Genes below the dotted line were unique to *A. thaliana*

SUPPLEMENTARY MATERIALS

Supplementary Tables, Datasets, and Figures for Chapter 2 are included with the electronic version of the dissertation.

REFERENCES

REFERENCES

Alexa, A. and Rahnenfuhrer, J. (2019). topGO: Enrichment analysis for gene ontology.

Bell, C.D., Soltis, D.E., and Soltis, P.S. (2010). The age and diversification of the angiosperms re-revisited. American Journal of Botany **97**: 1296–1303.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. **34**: 525–527.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M., and Buell, C.R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics **7**: 327.

Carović-Stanko, K., Liber, Z., Besendorfer, V., Javornik, B., Bohanec, B., Kolak, I., and Satovic, Z. (2010). Genetic relations among basil taxa (*Ocimum* L.) based on molecular markers, nuclear DNA content, and chromosome number. Plant Systematics and Evolution **285**: 13–22.

Carpenter, Eric J., Naim Matasci, Saravanaraj Ayyampalayam, Shuangxiu Wu, Jing Sun, Jun Yu, Fabio Rocha Jimenez Vieira, et al. (2019). Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). GigaScience **8**.

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J. **66**: 212–229.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud□Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. **89**: 789–804.

Crocoll, C., Asbach, J., Novak, J., Gershenzon, J., and Degenhardt, J. (2010). Terpene synthases of oregano (*Origanum vulgare* L.) and their roles in the pathway and regulation of terpene biosynthesis. Plant Mol. Biol. **73**: 587–603.

Croteau, R., Kutchan, T.M., and Lewis, N.G. (2000). Natural Products (Secondary Metabolites). In Biochemistry and Molecular Biology of Plants, B.B. Buchanon, W. Gruissem, and R.L. Jones, eds (American Society of Plant Physiologists).

Daferera, D.J., Tarantilis, P.A., and Polissiou, M.G. (2002). Characterization of essential oils from Lamiaceae species by Fourier Transform Raman Spectroscopy. Journal of Agricultural and Food Chemistry **50**: 5503–5507.

Daferera, D.J., Ziogas, B.N., and Polissiou, M.G. (2000). GC-MS analysis of essential oils from some Greek aromatic plants and their fungitoxicity on *Penicillium digitatum*. Journal of Agricultural and Food Chemistry **48**: 2576–2581.

Davidson, R. M., Hansey, C.N, Gowda, M., Childs, K.L., Lin, H. Vaillancourt, B. Sekhon, R. et al. 2011. Utility of RNA sequencing analysis of maize reproductive transcriptomes. The Plant Genome 4: 191–203.

Dong, A.X., Xin, H.B., Li, Z.J., Liu, H., Sun, Y.Q., Nie, S., Zhao, Z.N., Cui, R.F., Zhang, R.G., Yun, Q.Z., et al. (2018). High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. Gigascience **7**: giy068. doi: 10.1093/gigascience/giy068

Dudai, N., Carp, M.-J., Milavski, R., Chaimovitsh, D., Shachter, A., Baruch, K., Ronen, G., Gonda, I. (2018) High-quality assembly of sweet basil genome. bioRxiv: doi.org/10.1101/476044.

Dudareva, N., Andersson, S., Orlova, I., Gatto, N., Reichelt, M., Rhodes, D., Boland, W., and Gershenzon, J. (2005). The nonmevalonate pathway supports both monoterpene and sesquiterpene formation in snapdragon flowers. Proc. Natl. Acad. Sci. U. S. A. **102**: 933–938.

El-Gazzar, A. and Watson, L. (1970). A Taxonomic Study of Labiatae and Related Genera. New Phytol. **69**: 451–486.

El-Gebali, S., Mistry, J,. Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2018). The Pfam protein families database in 2019. Nucleic Acids Research **47**: D427–D432.

Emms, D.M. and Kelly, S. (2018a). OrthoFinder2: Phylogenomic orthology inference for comparative genomics. Genome Biology 20:238.

Emms, D.M. and Kelly, S. (2018b). STAG: Species Tree Inference from All Genes. bioRxiv. doi.org/10.1101/267914

Emms, D.M. and Kelly, S. (2017). STRIDE: Species Tree Root Inference from Gene Duplication Events. Mol. Biol. Evol. **34**: 3267–3278.

Fischer, N., Nitz, S., and Drawert, F. (1988). Original composition of marjoram flavor and its changes during processing. Journal of Agricultural and Food Chemistry **36**: 996–1003.

Fischer, N., Nitz, S., and Drawert, F. (1987). Original flavour compounds and the essential oil composition of marjoram (*Majorana hortensis* Moench). Flavour and Fragrance Journal **2**: 55–61.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**: 644–652.

Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using

CAFE 3. Mol. Biol. Evol. **30**: 1987–1997.

Harley RM, Atkins S, Budantsev AL, Cantino PD, Conn BJ, Grayer R, Harley MM, de Kok R, Krestovskaja T, Morales R, et al (2004) Labiatae. Flowering Plants · Dicotyledons. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 167–275

Helmkampf, M., Bellinger, M.R., Geib, S.M., Sim, S.B., and Takabayashi, M. (2019). Draft genome of the rice coral *Montipora capitata* obtained from linked-read sequencing. Genome Biol Evo **11**: 2045–2054.

Hulse-Kemp, A.M., Maheshwari, S., Stoffel, K., Hill, T.A., Jaffe, D., Williams, S.R., Weisenfeld, N., Ramakrishnan, S., Kumar, V., Shah, P., et al. (2018). Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. Hortic Res **5**: 4.

Ietswaart, J.H. (1980). A taxonomic revision of the genus *Origanum* (Labiatae) (Leiden University Press).

Iijima, Y., Davidovich-Rikanati, R., Fridman, E., Gang, D.R., Bar, E., Lewinsohn, E., and Pichersky, E. (2004). The biochemical and molecular basis for the divergent patterns in the biosynthesis of terpenes and phenylpropenes in the peltate glands of three cultivars of basil. Plant Physiology **136**: 3724 LP – 3736.

Jamshidi, R., Afzali, Z., and Afzali, D. (2009). Chemical composition of hydrodistillation essential oil of rosemary in different origins in Iran and comparison with other countries. American-Eurasian Journal of Agricultural and Environmental Sciences **5**: 78–81.

Jedrzejczyk, I. (2018). Study on genetic diversity between *Origanum* L. species based on genome size and ISSR markers. Industrial Crops and Products **126**: 201–207.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics **30**: 1236–1240

Kautsar, S.A., Suarez Duran, H.G., and Medema, M.H. (2018). Genomic identification and analysis of specialized metabolite biosynthetic gene clusters in plants using PlantiSMASH. In Plant Chemical Genomics: Methods and Protocols, F. Fauser and M. Jonikas, eds (Springer New York: New York, NY), pp. 173–188.

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

Köllner, T.G., Held, M., Lenk, C., Hiltpold, I., Turlings, T.C.J., Gershenzon, J., and Degenhardt, J. (2008). A maize (E)-beta-caryophyllene synthase implicated in indirect defense responses against herbivores is not expressed in most American maize varieties. The Plant cell **20**: 482–494.

Köllner, T.G., Schnee, C., Gershenzon, J., and Degenhardt, J. (2004). The variability of sesquiterpenes emitted from two *Zea mays* cultivars is controlled by allelic variation of two terpene synthase genes encoding stereoselective multiple product enzymes. The Plant Cell **16**: 1115 LP – 1131.

Komaitis, M.E., Ifanti-Papatragianni, N., and Melissari-Panagiotou, E. (1992). Composition of the essential oil of marjoram (*Origanum majorana* L.). Food Chemistry **45**: 117–118.

Koroch, A.R., Wang, W., Michael, T.P., Dudai, N., Simon, J.E., and Belanger, F.C. (2010). Estimation of nuclear DNA content of cultivated *Ocimum* species by using flow cytometry. Israel Journal of Plant Sciences **58**: 183–189.

Kyriakidou, M., Anglin, N.L., Ellis, D., Tai, H.H., and Strömvik, M.V. (2020). Genome assembly of six polyploid potato genomes. Sci Data **7**: 88.

Lau, K.H., Bhat, W.W., Hamilton, J.P., Wood, J.C., Vaillancourt, B., Wiegert-Rininger, K., Newton, L., Hamberger, B., Holmes, D., Hamberger, B., and Buell, C.R. (2020). Genome assembly of *Chiococca alba* uncovers key enzymes involved in the biosynthesis of unusual terpenoids. DNA Res. **27**: dsaa013, https://doi.org/10.1093/dnares/dsaa013

Leebens-Mack, J.H. et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**: 679–685.

Li, B., Cantino, P.D., Olmstead, R.G., Bramley, G.L., Xiang, C.L., Ma, Z.H., Tan, Y.H., and Zhang, D.X. (2016). A large-scale chloroplast phylogeny of the Lamiaceae sheds new light on its subfamilial classification. Sci. Rep. **6**: 34343.

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

Li, P., Qi, Z.-C., Liu, L.-X., Ohi-Toma, T., Lee, J., Hsieh, T.-H., Fu, C.-X., Cameron, K.M., and Qiu, Y.-X. (2017). Molecular phylogenetics and biogeography of the mint tribe Elsholtzieae (Nepetoideae, Lamiaceae), with an emphasis on its diversification in East Asia. Scientific Reports **7**: 2057.

Lichman, B.R., Godden, G.T., Hamilton, J.P., Palmer, L., Kamileen, M.O., Zhao, D., Vaillancourt, B., Wood, J.C., Sun, M., Kinser, T.J., et al. (2020a). The evolutionary origins of the cat attractant nepetalactone in catnip. Science Advances **6**: eaba0721.

Lichman, B.R., Godden, G.T., and Buell, C.R. (2020b). Gene and genome duplications in the evolution of chemodiversity: perspectives from studies of Lamiaceae. Curr. Opin. Plant Biol. **55**: 74–83.

Lukas, B., Schmiderer, C., and Novak, J. (2015). Essential oil diversity of European *Origanum vulgare* L. (Lamiaceae). Phytochemistry **119**: 32–40.

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L.L., and Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant

phylogenetic diversity. New Phytologist **207**: 437–453.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**: 764–770.

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

Mint Evolutionary Genetics Consortium (2018). Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. Molecular Plant **11**: 1084–1096.

Mowforth MA (1985) Variation in nuclear DNA amounts in flowering plants: an ecological analysis. PhD thesis. University of Sheffield

Nützmann, H.-W., Huang, A., and Osbourn, A. (2016). Plant metabolic clusters – from genetics to genomics. New Phytologist **211**: 771–789.

Ott, A., Schnable, J.C., Yeh, C.-T., Wu, L., Liu, C., Hu, H.-C., Dalgard, C.L., Sarkar, S., and Schnable, P.S. (2018). Linked read technology for assembling large complex and polyploid genomes. BMC Genomics **19**: 651.

Ozerov, M.Y., Ahmad, F., Gross, R., Pukk, L., Kahar, S., Kisand, V., and Vasemägi, A. (2018). Highly continuous genome assembly of Eurasian perch (*Perca fluviatilis*) using linked-read sequencing. G3: Genes|Genomes|Genetics **8**: 3737 LP – 3743.

Pellicer, J., Estiarte, M., Garcia, S., Garnatje, T., Peñuelas, J., Sardans, J., and Vallès, J. (2010). Genome size unaffected by moderate changes in climate and phosphorus availability in Mediterranean plants. African Journal of Biotechnology **9**: 6070–6077.

Rastogi, S., Meena, S., Bhattacharya, A., Ghosh, S., Shukla, R.K., Sangwan, N.S., Lal, R.K., Gupta, M.M., Lavania, U.C., Gupta, V., Nagegowda, D.A., and Shasany, A.K. (2014). De novo sequencing and comparative analysis of holy and sweet basil transcriptomes. BMC Genomics **15**: 588.

Rewers, M. and Jedrzejczyk, I. (2016). Genetic characterization of *Ocimum* genus using flow cytometry and inter-simple sequence repeat markers. Industrial Crops and Products **91**: 142–151.

Sasikumar, B. (2012). Rosemary. In Handbook of Herbs and Spices, K.V. Peter, ed (Woodhead), pp. 452–468.

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

Simon JE, Morales MR, Phippen WB, Vieira RF, Hao Z (1999) Basil: a source of aroma compounds and a popular culinary and ornamental herb. *In* J Janick, ed, Perspectives on

new crops and new uses. American Society for Horticultural Science, Alexandria, VA, pp 499–505

Sivropoulou, A., Papanikolaou, E., Nikolaou, C., Kokkini, S., Lanaras, T., and Arsenakis, M. (1996). Antimicrobial and Cytotoxic Activities of Origanum Essential Oils. Journal of Agricultural and Food Chemistry **44**: 1202–1205.

Skoula, M., Gotsiou, P., Naxakis, G., and Johnson, C.B. (1999). A chemosystematic investigation on the mono- and sesquiterpenoids in the genus *Origanum* (Labiatae). Phytochemistry **52**: 649–657.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**: 637–644.

Sun, H., Ding, J., Piednoël, M., and Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. Bioinformatics **34**: 550–557.

Torre, S., Tattini, M., Brunetti, C., Guidi, L., Gori, A., Marzano, C., Landi, M., and Sebastiani, F. (2016). De novo assembly and comparative transcriptome analyses of red and green morphs of sweet basil grown in full sunlight. PLOS ONE **11**: e0160370.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**: 511–515.

Turner, G.W. and Croteau, R. (2004). Organization of monoterpene biosynthesis in *Mentha*. Immunocytochemical localizations of geranyl diphosphate synthase, limonene-6-hydroxylase, isopiperitenol dehydrogenase, and pulegone reductase. Plant Physiol. **136**: 4215–4227.

Walker, J.B., Sytsma, K.J., Treutlein, J., and Wink, M. (2004). *Salvia* (Lamiaceae) is not monophyletic: implications for the systematics, radiation, and ecological specializations of Salvia and tribe Mentheae. American Journal of Botany **91**: 1115–1125.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. **35**: 543–548.

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. Genome Res. **27**: 757–767.

Workman, R., Fedak, R., Kilburn, D., Hao, S., Liu, K., et al (2018) High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. Nature Protocol Exchange. doi: 10.1038/protex.2018.059

Zaouali, Y., Bouzaine, T., and Boussaid, M. (2010). Essential oils composition in two *Rosmarinus officinalis* L. varieties and incidence for antimicrobial and antioxidant activities. Food and chemical toxicology : An international journal published for the British Industrial Biological Research Association **48**: 3144–3152.

Zeng, L., Zhang, N., Zhang, Q., Endress, P.K., Huang, J., and Ma, H. (2017). Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. New Phytologist **214**: 1338–1354.

Zhao, D., Hamilton, J.P., Bhat, W.W., Johnson, S.R., Godden, G.T., Kinser, T.J., Boachon, B., Dudareva, N., Soltis, D.E., Soltis, P.S., Hamberger, B., and Buell, C.R. (2019). A chromosomal-scale genome assembly of *Tectona grandis* reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. Gigascience **8(3)**: giz005.

CHAPTER 3

GENOMIC VARIATION WITHIN THE MAIZE STIFF STALK HETEROTIC GERMPLASM POOL

This chapter was published in the following manuscript:

Bornowski N, Michel KJ, Hamilton JP, Ou S, Seetharam AS, Jenkins J, Grimwood J, Plott C, Shu S, Talag J, et al (2021) Genomic variation within the maize stiff-stalk heterotic germplasm pool. *Plant Genome* 14: e20114. doi: 10.1002/tpg2.20114.

ABSTRACT

The Stiff Stalk heterotic group is an important source of inbreds used in U.S. commercial hybrid production. Founder inbreds B14, B37, B73, and to a lesser extent B84, are found in the pedigrees of a majority of commercial seed parent inbred lines. We created high-quality genome assemblies of B84 and four ex-Plant Variety Protection lines LH145 representing B14, NKH8431 of mixed descent, PHB47 representing B37, and PHJ40 which is a Pioneer Hi-Bred early Stiff Stalk type. Sequence was generated using long-read sequencing achieving highly contiguous assemblies of 2.13 to 2.18 Gbp with N50 scaffold lengths greater than 200 Mbp. Inbred-specific gene annotations were generated using a core five-tissue gene expression atlas while transposable element annotation was conducted using *de novo* and homology-directed methodologies. In comparison to the reference inbred B73, synteny analyses revealed extensive collinearity across the five Stiff Stalk genomes, although unique components of the maize pan-genome were detected. Comparison of this set of Stiff Stalk inbreds with the original Iowa Stiff Stalk Synthetic breeding population revealed that these inbreds represent only a proportion of variation in the original Stiff Stalk pool and there are highly conserved haplotypes in released public and ex-Plant Variety Protection inbreds. Despite the reduction in variation from the original Stiff Stalk population, substantial genetic and genomic variation was identified supporting the potential for continued breeding success in this pool. The assemblies described here represent Stiff Stalk inbreds that have historical and commercial relevance and provide further insight into the emerging maize pan-genome.

INTRODUCTION

Maize production is vital to American agriculture and the global food supply, and significant heterosis, or the superior performance of a hybrid progeny over its inbred parents, exists in maize. Heterosis generated from the cross of two unrelated inbreds from opposing heterotic groups has supported immense yield gains since the introduction of the hybrid cross in the early 20th century. Modern maize breeding relies on several key heterotic groups and subgroups (White et al., 2020) with new inbreds generated within heterotic groups and hybrids generated from crosses between heterotic groups. Heterotic patterns did not arise out of a conscious decision to create them, but rather as a necessity for organization within breeding programs (Tracy and Chandler, 2006). Initial pools were made arbitrarily by some programs, while others attempted to group related lines together. For example, Pioneer Hi-Bred made efforts to gather good seed parents in one group and good pollen parents in the other (Tracy and Chandler, 2006). Over time, the contrasting pools genetically diverged, as evidenced by a study of inbreds used from the early 1930's to 2001 at Pioneer Hi-Bred (Duvick, 2005). Using simple sequence repeat markers and multidimensional scaling, the author demonstrated that inbreds used in the "pre-heterotic" era do not cluster in a discernible pattern, while advanced inbreds classified as either Stiff Stalk or Non-Stiff Stalk form two distinct groups (Duvick, 2005). This allelic diversity led to the great success of the heterotic pattern breeding method as alleles are fixed for contrasting allelic states between heterotic pools, contributing to additive-by-additive epistasis and repulsion phase linkages that create pseudo-overdominance (Graham et al., 1997; Larièpe et al., 2012).

Corporations, individuals, and public institutions can protect inbred lines with Plant Variety Protection (PVP) certificates, which allow the breeder or organization sole ownership over sales of the hybrid progeny for 20 years in the case of maize, at which point, the certificates expire ([USC04] 7 USC Ch. 57: Plant Variety Protection, 1970)). The rapidly increasing number of expired PVP (ex-PVP) certificates gives public entities the unique opportunity to characterize the pedigrees, genetic diversity, and phenotypic characteristics of elite ex-PVP lines that originate from a diverse group of breeding programs and contain the parent inbreds that have supported the hybrid maize industry. Several heterotic groups have emerged over the last few decades, which can be studied as the PVP certificates on inbreds expire and biological materials become freely available. Broadly, the major groups are the Stiff Stalk, Iodent, and non-Stiff Stalk heterotic pools. The Iodent group as represented in ex-PVP inbreds was founded by PH207 (Hirsch et al., 2016), and has the most limited genetic diversity. The Stiff Stalk heterotic pool, as described below, is also more limited in diversity than the non-Stiff Stalk pool which comprises most other lines not grouping as Iodent or Stiff Stalk. Each group has a unique history of selection and development.

The Stiff Stalk heterotic group originated from the Iowa Stiff Stalk Synthetic (BSSS) developed by Dr. George Sprague at Iowa State University in the 1930's. BSSS is composed of 16 inbred lines primarily of Reid Yellow Dent heritage, and underwent several cycles of recurrent selection (Troyer, 1999). The population has yielded several key founder inbreds, including B14, released by Sprague in 1953, B37, released by Sprague in 1958, and B73, released by Dr. Wilbert Russell in 1972 (Troyer, 1999). Related samples of the population were used in other public breeding programs and resulted in release of inbreds including N7A and

N28, for example (npgsweb.ars-grin.gov). B14 was a first cycle selection from the BSSS and was chosen for its superior yield, stalk and root strength, and was used heavily in the development of inbreds adapted for early maturity zones such as the northern United States, Canada, and Europe including A632 and A634 (Troyer, 1999). B37 was also released from the first cycle of selection of the BSSS due to its positive contributions to hybrid yield and agronomic quality but faced issues of low pollen shed and a protracted anthesis-silking interval (Troyer, 1999). B73 was chosen from cycle five for its high yield in test cross hybrids (Troyer, 1999). B73 x Mo17 was an incredibly popular hybrid grown across the American corn belt during the 1970's, and B73 would later serve as the first representative reference assembly of maize (Schnable et al., 2009). Goodman (1990) estimated that perhaps 70% of the hybrid commercial germplasm in 1990 relied on close relatives of just six inbreds including Lancaster lines C103, Mo17, and Oh43, and Stiff Stalk lines B73, B37, and A632 (a B14 derivative) as the seed parent. These three Stiff Stalk inbreds were heavily utilized by private seed companies as foundational inbreds within their breeding programs and were valued for their superior seed parent characteristics. Thus, the Stiff Stalk heterotic group was, and is, vital for North American hybrid maize production.

The first maize reference genome assembly was generated from B73 (Schnable et al., 2009). Several maize genome assemblies have since been published including tropical lines CML247 and SK (Lu et al., 2015; Yang et al., 2019), Iodent line PH207 (Hirsch et al., 2016), Lancaster line Mo17 (Sun et al., 2018), W22 (Springer et al., 2018), European Flint lines EP1, F7, DK105, and PE0075 (Haberer et al., 2020), Oh43-type line PHJ89 (Gage et al., 2019), sweet corn line Ia453-*sh2* (Hu et al., 2021) and teosinte *Zea mays* ssp. *mexicana* (Yang et al., 2017b).

Structural variation, including copy number variants (CNV) and their more extreme structural variant, presence absence variants (PAV), have been documented in maize and are known to influence phenotypes in a number of crop and model species (Cook et al., 2012; Qi et al., 2013; Chang et al., 2015; Hardigan et al., 2016; Gordon et al., 2017; Hardigan et al., 2017; Ou et al., 2018a; Wang et al., 2018; Gao et al., 2019; Pucker et al., 2019; Song et al., 2020). Abundant gene content variation exists between the commercial inbreds B73 and PH207 (Hirsch et al., 2016), though syntenic genes are highly conserved between the two lines and differential fractionation plays a limited role in generating gene content variation (Brohammer et al., 2018). Contributions from *Z. mays* ssp. *mexicana* have contributed to modern maize adaptation and improvement (Yang et al., 2017b), and comparisons between W22 and B73 demonstrated CNV of transposable elements (TEs), which influence the study of functional genomics and the impact of TEs on complex phenotypes (Springer et al., 2018). As a result of this pan-genome level variation, candidate gene predictions can depend on the reference line used for calling single nucleotide polymorphisms (SNPs), and structural variation between reference lines can influence genome-wide association study results (Gage et al., 2019).

To better understand the genomic diversity present within this important commercial germplasm group, and to support ongoing genetic and functional studies, five inbreds that represent the diversity and history of the Stiff Stalk heterotic group (Table 3.1) were sequenced. All pedigree and accession information was compiled from the Germplasm Resource Information Network Database (npgsweb.ars-grin.gov). B84 was released in 1979 as a cycle seven selection of the BSSS with *Helminthosporium turcicum* resistance (BSSS(HT)C7, now known as *Setosphaeria turcica*, common name Northern Corn Leaf Blight). LH145 is a

derivative of B14 through both parents, A632Ht and CM105, and was protected by Holden's Foundation Seed, Inc. in 1984. NKH8431, also known as H8431, was developed by Northrup, King and Company, was protected in 1988, and was the result of a cross between B73-like and B14-like proprietary lines. PHB47 was protected via a PVP certificate by Pioneer Hi-Bred, International (hereafter PHI) in 1984, and was the result of crossing B37 with SD105, an early inbred developed by South Dakota State University. During PHB47's development, populations were backcrossed twice to B37. Finally, PHJ40 is the earliest flowering of the group, was developed by PHI by crossing proprietary inbred lines, and was protected by PVP certificate in 1987. While the subheterotic groups of the parents of PHJ40 are not known, previous work has shown PHJ40 has admixture derived membership with the B37 subgroup (White et al., 2020).

MATERIALS AND METHODS

Genome Sequencing and Assembly

DNA Isolation

High molecular weight DNA was extracted from young leaves using the protocol of Doyle and Doyle (1987) with minor modifications. In brief, young leaves were flash frozen and ground to a fine powder in a frozen mortar with liquid N_2 followed by very gentle extraction for 1hr at 50 °C in cetyl trimethylammonium bromide buffer that included proteinase K, polyvinylpyrrolidone-40, and beta-mercaptoethanol. After centrifugation, the supernatant was gently extracted twice with 24:1 chloroform:iso-amyl alcohol. The upper phase was adjusted to $1/10^{th}$ volume with 3M potassium acetate, gently mixed, and the DNA was precipitated with iso-propanol. DNA was collected by centrifugation, washed with 70% ethanol, air dried for 20 min and dissolved thoroughly in 1x 10 mM Tris-Cl, 1 mM ethylene diaminetetraacetic acid at room temperature; DNA size was validated by pulsed field gel electrophoresis.

Genome Sequencing

Zea mays inbreds (B84, LH145, NKH8431, PHB47, PHJ40) were sequenced using a whole genome shotgun sequencing strategy. Sequencing reads were generated using Illumina HiSeq-2500 and PacBio Sequel I platforms (Supplementary Table S3.1) at the Department of Energy Joint Genome Institute and the HudsonAlpha Institute. For the PacBio sequencing, an average of 50.8 chips per variety were collected (10-hour movie time) that yielded 88.4x, 112.2x, 113.7x, 71.2x, and 85.4x coverage for B84, LH145, NKH8431, PHB47, PHJ40, respectively. The Illumina read sets consist of 62.8x to 69.4x coverage of high-quality Illumina bases for each inbred.

Assembly and Integration

The genomes were assembled using the MECAT assembler (v1.2) (Xiao et al., 2017) and polished using ARROW (v2.2.2) (Chin et al., 2013). To identify false joins, 28,964 non-repetitive, non-redundant, 1,500 bp syntenic markers were extracted from the B73 v4 assembly and used to first resolve misjoins and then orient, order, and join the contigs into 10 chromosomes using the B73 markers. Telomeres were evaluated by searching for the kmer $(TTTAGGG)_n$ where the value of n varied from nine to 20; the longest run of telomere was identified for each contig containing a telomere and placed at the ends of the chromosomes. Remaining scaffolds were screened against the NR GenBank database to remove contamination. Homozygous SNPs and insertion-deletions (INDELs), representing remaining PacBio errors, were corrected using 60x of Illumina reads (2x150, 400bp insert) by aligning the reads using BWA-MEM (v0.7.15) (Li and Durbin, 2009) and identifying homozygous SNPs and INDELs with the GATK UnifiedGenotyper tool (v3.6) (McKenna et al., 2010). The final genome assemblies had 86.6% to 98.4% of the sequence anchored to the 10 chromosomes with N50 contig lengths ranging from 893.8 kbp to 3.1 Mbp.

Genome Quality Assessments

Whole Genome Shotgun Sequence Read Alignment

Whole genome shotgun (WGS) libraries from the five inbreds were aligned to their cognate genome assemblies (Supplementary Table S3.2) to assess the quality of the assemblies. Read quality was inspected with FastQC v0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) before processing with Cutadapt v1.18 (Martin, 2011) to remove sequencing adapters and low-quality reads using the parameters

"-q 10 -n 2 -m 31". B73 WGS reads were clipped to 150 nt using the Cutadapt parameters "-u 7 -u 93 and -U 7 -U 93" prior to adapter trimming. Additionally, processed B73 WGS reads were randomly subsampled with the reformat.sh script from the BBMap suite v37.61 (https://sourceforge.net/projects/bbmap/) using "sampleseed=100" to obtain similar read quantities as the other libraries. Cutadapt-filtered WGS reads were aligned to their cognate genome assembly using BWA-MEM v0.7.16a (Li and Durbin, 2009) with the "-M" flag used to mark shorter split hits as secondary, "-t" specifying 22 threads, and "-R" specifying read group headers.

RNA-sequencing Read Alignment

Illumina RNA-sequencing (RNA-seq) libraries from internode, shoot, leaf, root, and endosperm tissue from each inbred (Li et al., 2020) were used for genome annotation and estimation of expression abundance. Read quality was inspected with FastQC v0.11.8 before processing with Cutadapt v1.18 (Martin, 2011) to remove sequencing adapters and low-quality reads using the parameters "-q 10 -n 2 -m 31". Cutadapt-filtered RNA-Seq reads were aligned to their cognate genome assembly using the splice-site aware algorithm HISAT2 v2.2.0 (Kim et al., 2019) in RF stranded mode with parameters "--max-intronlen 12000 bp, --dta-cufflinks, --no-unal, --no-summary".

Benchmarking Universal Single Copy Orthologs (BUSCO)

Genome assemblies were queried for conserved single-copy orthologs using BUSCO (Simao et al., 2015; Waterhouse et al., 2018) to assess genic completeness. Additionally, genome assemblies of maize lines B73 v4 (downloaded from ftp://ftp.gramene.org/pub/gramene/release-

59/fasta/zea_mays/), PH207 (downloaded from https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=ZmaysPH207), Mo17 (downloaded from https://download.maizegdb.org/Zm-Mo17-REFERENCE-CAU-1.0/), and Ia453-*sh2* (downloaded from https://www.ncbi.nlm.nih.gov/assembly/GCA_016432965.1) were also queried. BUSCO v4.1.4 was run in genome mode using the Embryophyta odb10 dataset (creation date: 2019-11-20, number of species: 50, number of BUSCOs: 1614) with default parameters.

Long Terminal Repeat Assembly Index (LAI)

The assembly contiguity of the TE space of each genome was evaluated using LAI (beta 3.2) (Ou et al., 2018b) from the LTR_retriever (v2.9.0) package (Ou and Jiang, 2018) with parameters "-intact file4 -all file3 -q -totLTR 76.34 -iden 94.854 -t 10". The "-intact" file was generated using EDTA (v1.9.0) (Ou et al., 2019) as described below. The "-all" file was the RepeatMasker out file of each genome annotated by the pan-Stiff Stalk TE library (see 2.3.2 for details on generation of the library).

Genome Annotation

Construction of the Pan-Stiff Stalk Transposable Element Library

A manually curated Transposable Element library from the Maize TE Consortium (Schnable et al., 2009) (MTEC, downloaded from https://github.com/oushujun/MTEC) was used as the base library, and supplemented with novel TE families identified from the six genomes, including the five Stiff Stalk genomes reported in this study and the B73 v4 genome. The EDTA package (v1.9.0) (Ou et al., 2019) was used to identify novel TEs of each genome with

parameters “--cds” and “--curatedlib”. With the “--cds” parameter, coding sequences annotated from each genome were provided to remove gene-related sequences in the resulting TE library. With the “--curatedlib” parameter, the base library (i.e., the MTEC library) was provided for EDTA to identify novel TE families beyond those already present in the MTEC library. The six novel TE libraries were combined with the MTEC library using the Perl script “make_panTElib.pl” in the EDTA package. The 80-95-80 rule (80% identity, 95% coverage, 80bp minimum length) was used to cluster redundant sequences with parameters “-miniden 80 -mincov 0.95 -minlen 80”.

Annotation of Pan-Genome TEs

Transposable element annotation of each genome was performed based on both structural- and homology-annotations using EDTA (v1.9.0) (Ou et al., 2019) and RepeatMasker (v4.0.9) (http://www.repeatmasker.org/). First, each genome was annotated using the pan-genome TE library and RepeatMasker with parameters “-q -no_is -norna -nolow -div 40”, allowing up to 40% sequence divergence. EDTA was executed again on the original structural annotation of each genome to unify TE family names, with parameters “--cds file1 --curatedlib file2 --step anno --rmout file3 ---anno 1 --evaluate 1”. The “--cds” file was the same coding sequences for each genome previously provided. The “--curatedlib” file was the pan-Stiff Stalk TE library. The “--rmout” file was the RepeatMasker out file of each genome annotated by the pan-Stiff Stalk TE library. The insertion time of each LTR retrotransposon was estimated by LTR_retriever (v2.9.0) (Ou et al., 2018b) with $T = K/2\mu$, where K is the divergence between the left and right LTR of the element and $\mu = 3.3e\text{-}8$ per bp per year for heterochromatic regions (Clark et al., 2005).

Annotation of Gene Models

Each of the six Stiff Stalk genomes, including B73, were annotated for gene models using an identical pipeline using inbred-specific transcript evidence thereby eliminating false annotations from transcripts from other inbreds. RNA-seq libraries were cleaned using Cutadapt (v2.9) (Martin, 2011) using the parameters "--times 2 --minimum-length 100 --quality-cutoff 10". Cleaned reads from each library were then aligned to their respective genome assembly using HISAT2 (v2.2.0) (Kim et al., 2019) with the parameters "--max-intronlen 5000 --rna-strandness RF --no-unal --dta", and assembled using Stringtie (v2.1.1) (Kovaka et al., 2019) with the parameter "--rf" and the assembled transcript sequences extracted with gffread (v0.11.7) (Pertea and Pertea, 2020).

Each genome assembly was masked with RepeatMasker (v4.1.0) (http://www.repeatmasker.org/) using the curated maize repeat library maizeTE02052020 (https://github.com/oushujun/MTEC) using the parameters "-e ncbi -s -nolow -no_is -gff". Augustus (v3.3.3) (Stanke et al., 2008) was used to generate gene predictions on the masked assemblies using the maize5 training parameter set and the RNA-Seq alignments as hints. The gene predictions were refined using PASA2 (v2.4.1) (Haas et al., 2005) (http://pasapipeline.github.io/) in two rounds of annotation comparison (-I 60000) using the RNA-Seq transcript assemblies as evidence to generate the working model gene set.

To identify high-confidence gene models, the working gene model set was searched against the PFAM database (v32) (Finn et al., 2016) with hmmscan (HMMER, v3.2.1) (Mistry et al., 2013) with a cutoff of "--domE 1e-3 -E 1e-5" to identify gene models encoding a Pfam

domain as described previously (Pham et al., 2020). Gene expression abundances for the working gene models (transcripts per million; (TPM)) were generated for each RNA-seq library using Kallisto (v0.46.0) (Bray et al., 2016). High confidence gene models were identified if they had a TPM value > 0 in at least one RNA-seq library and/or had a PFAM domain match. Partial gene models or gene models with matches to transposable element-related PFAM domains were also excluded from the high-confidence model set.

Functional annotation was assigned to the working gene model set using search results from the predicted proteins against the Arabidopsis proteome (TAIR10; Arabidopsis.org), the PFAM database (v32) (Finn et al., 2016), and Swiss-Prot plant proteins (release 2015_08). Results were processed in the same order (TAIR, PFAM, Swiss-Prot) and the function of the first informative hit was transitively assigned to the gene model.

Comparative Genome Analyses

Transcript Alignment

Annotated high confidence coding sequences (CDS) from all six genomes (B73, B84, LH145, NKH8431, PHB47, and PHJ40) were aligned to all six genome assemblies using GMAP (v20170905) (Wu and Watanabe, 2005) with thresholds of 95% identity and 95% coverage used to determine gene presence/absence. Sequences were considered present in a genome assembly if they aligned to either a unique location or multiple locations.

Structural Variation

Structural variants (SV) for the Stiff Stalk genomes were characterized as described previously (Hufford et al., 2021). Briefly, this SV-detection pipeline includes a combination of three different methods, using three different data types mapped against the B73 v4 reference (Jiao et al., 2017). The first approach involved mapping long reads from each genome to B73 v4, the second, aligning the chromosomal genome assemblies of each line to B73 v4, and the third, taking *in silico* digested assemblies (to simulate a BioNano optical map) of each maize line and aligning these to the simulated B73 optical map. These approaches were used to characterize SVs separately and then collapsed to generate a comprehensive set of SVs for the five Stiff Stalks.

Error corrected long reads of each Stiff Stalk maize inbred were mapped to the B73 v4 genome using a sensitive mapping program, NGMLR (v0.2.7) (Sedlazeck et al., 2018). All options were set to default, except for the "--presets" option, which was set to "pacbio", and the "--bam-fix" option, which enables bam compatible output files. In order to accelerate the mapping step, input files (PacBio reads) were split into smaller subsets, and mapping was performed in parallel to the reference genome, followed by concatenation of bam files to a single file using samtools merge (v1.9) (Li et al., 2009). The final BAM file for each maize line was then used with SNIFFLES (v1.0.11) (Sedlazeck et al., 2018) in order to call structural variants in two iterations. For the first iteration, SNIFFLES was run using stringent parameters "--max_num_splits 2, --min_support 20, --min_zmw 2, --min_seq_size 5000, --max_distance 5000, --cluster, and --cluster_support 2" with minimum SV size set to 100 bp "--min_length 100" and a VCF-formatted file generated for each maize inbred line. SURVIVOR (v1.0.6) (Jeffares et al.,

2017) was then used to merge individual VCF files, with options max distance between breakpoints set to 1000 and taking SV type and strand into account. We did not use the options to estimate SV size nor to take the minimum size of SV into account in order to generate a joint SV VCF file. Missing and absent SV calls across lines were filled in a second iteration of SNIFFLES. For this run, the merged SVs were provided as input (--Ivcf) along with the original BAM files (mapped reads). The final genotyped SVs were then once again combined using SURVIVOR and filters were applied to limit SVs to a size range of 100 bp to 100 kbp.

Each of the Stiff Stalk inbred assemblies was aligned against the B73 v4 reference, using minimap2 (v2.17-r941) (Li, 2018) to generate PAF-formatted alignment files (default options with -c to enable cigar strings in the output files, -x asm5 to use a ~0.1% sequence divergence preset and --cs to encode bases at mismatches and the INDELs options). The PAF files were sorted using the UNIX sort command, and INDELS were inferred using paftools (k8 paftools.js call) (Li, 2018). The native tab-separated output files were converted to BED format using awk in order to visualize INDELS and syntenic blocks in the IGV genome browser (Robinson et al., 2011).

For larger SVs (>100 kbp) that could not be characterized using long reads or aligned using genome-to-genome-based alignment methods, we used optical-map-based SV detection. In this approach, the maize genome was first subjected to *in silico* digestion using the fa2cmap_multi_color.pl script in the BioNano solve program and the CTTAAG enzyme motif in order to simulate a contiguous Bionano optical map for each chromosome. Second, CMAP format BioNano maps were aligned against the B73 CMAP file using the RefAligner tool from

runCharacterize.py and the runSV.py script from BioNano solve. Since labelled markers are aligned instead of individual bases, accurate detection of large-scale inversions, deletions and insertions can be achieved; however, smaller SVs are difficult to detect. Default options were used for both steps, with the arguments supplied through an XML file (optArguments_nonhaplotype_noES_DLE1_saphyr.xml). In the third step, the resulting smap file from the second step (with the list of structural variants detected between query maps and reference maps in tsv format), was converted to VCF formatted files using the smap_to_vcf_v2.py script. The final SV file in VCF format was filtered to only include SVs greater than 100 kbp using an awk command. BioNano-based SV identification was carried out using two different enzymes and the breakpoints were manually inspected using bed files generated from genome-to-genome alignments in IGV and synteny dot-plots before finalizing the SV calls. The BioNano SV start and stop sites were refined based on the consensus positions determined by two enzymes independently along with genome-to genome alignments. The final curated SVs were merged to generate a joint SV file using SURVIVOR, with similar options as detailed above. The final SV set was generated by merging the SNIFFLES SVs with the curated BioNano SVs.

To characterize SVs within the 9 to 11 kbp size class, which was enriched in deletions, we annotated the deleted sequence from B73 with the pan-genome TE library using RepeatMasker and assessed enrichment for fl-LTRs, which typically fall in this size range, using the 80-80-80 rule that required at least 80 bp, 80% identity, and 80% coverage for the matching LTR sequence. We also extracted random sequences mimicking the exact length of these deletions in the B73v4 genome and performed the same annotation with 10 iterations.

Syntenic Analysis of Gene Content across the Inbreds

Syntenic regions among the six Stiff Stalk inbreds were identified using the MCScanX (v20170322) toolkit (Wang et al., 2012). The MCScanX algorithm was run with default parameters on each inbred using B73 v4 MSU annotation generated in this study as the reference to determine collinear blocks of genes.

Orthology and Paralogy Analysis

Orthologous and paralogous genes among the six Stiff Stalk genomes were identified using Orthofinder (v2.5.1) (Emms and Kelly, 2019). Analyses were conducted using the predicted proteomes from each Stiff Stalk genome with default settings. Orthologous groups represented by all inbreds were used to construct and root a consensus tree with the STAG and STRIDE algorithms, respectively (Emms and Kelly, 2017, 2018).

Resistance Gene Classification

Putative resistance genes were identified by querying high confidence representative peptides against the curated Pathogen Receptor Genes database (PRGdb) (v3.0) using the DRAGO2 API (Osuna-Cruz et al., 2018).

Identification of Descendant Regions

Single nucleotide polymorphisms (SNPs) generated using RNA-seq data with imputation of the 942 accessions in the Wisconsin Diversity panel (WiDiv-942) (Mazaheri et al., 2019), were used to generate haplotypes using the TASSEL 5.0 plugin FILLINFindHaplotypesPlugin (Bradbury et al., 2007). Default parameters were used except for the parameters "-mxDiv 0.03, -

minTaxa 1, -hapSize 1000, -minPres 250, and -extOut true". Thus, maximum divergence from the founder was set to 3%, the minimum number of taxa set to one to allow haplotypes found in a single individual, and the haplotype size was set to 960 SNP windows, as 960 is the closest multiple of 64 less than 1000. Haplotype data was processed and assigned to a hierarchy using the convert_fillinhaps_to_feather_or_csv.R and apply_hierarchy.R scripts (Coffman et al., 2020). This script names a representative inbred for each haplotype group based on a hierarchy, such that the highest ranking inbred within each group is listed as the representative. Ranking inbreds using a hierarchy allows for more convenient visualization of shared haplotype blocks and transmission through time and selection.

The WiDiv panel contains 15 inbreds that represent 13 of the original 16 BSSS founders, in addition to the parents of one of the unavailable lines. Inbreds Ind-461-3 and CI617 were not available, while inbreds B2 and Fe were included as the parents of unavailable inbred F1B1 (Gerke et al., 2015). A group of 41 unselected inbreds from the base BSSS population, hereafter BSSSC0, followed in the hierarchy. Previous work identified within the WiDiv 16 public inbreds that were classified according to pedigree information as directly selected from the BSSS germplasm and 21 ex-PVP inbreds that were derived from the Stiff Stalk founders B14, B37 and B73 according to ADMIXTURE analysis (Gage et al., 2019). These lines followed the BSSSC0 inbreds in our hierarchy, followed by any other remaining inbreds that clustered with Stiff Stalk founders B14, B37 and B73 according to ADMIXTURE analysis (Mazaheri et al., 2019). Lines were placed in order of year of release when that information was available, otherwise, lines were placed in the hierarchy in alphanumeric order within their groups. In addition, any haplotype groups that were represented by non-founder or BSSSC0 lines were set to be plotted in

white so that only haplotypes that were present in the base BSSS population would be plotted in color in the publicly released and ex-PVP lines. Once the hierarchy was constructed, a neighbor joining tree was made using default parameters in TASSEL 5.0 to order the inbreds along the x-axis according to genetic distance to facilitate visualization of shared haplotypes (Bradbury et al., 2007).

Allele frequencies were calculated for the base, unselected population, consisting of the founder and BSSSC0 lines, and for the selected population, consisting of the public and ex-PVP lines identified previously (Gage et al., 2019). F_{st} was calculated using vectorFst.R (Beissinger et al., 2014) available at http://beissingerlab.github.io/docs/vectorFst.R, including a correction for the small number of populations (Weir et al., 1984). SNPs were binned into the same windows used for haplotype analysis, and the window took on the maximum F_{st} value of SNPs within the window. Windows in the top 10^{th} percentile of genome-wide values were plotted in black alongside haplotypes for visualization.

Finally, SNP based identity by state (IBS) was calculated using the WiDiv-942 RNA-sequencing SNPs for the five inbreds compared to their most related founder Stiff Stalk lines. Values were averaged into bins using the same physical position boundaries as the previous haplotype plots. Approximate centromere locations, as determined by the mean physical position of the centromere in the maize B73-Ab10 assembly, are marked by vertical lines on each chromosome (Liu et al., 2020). Windows were noted as conserved if the average IBS was greater than 0.97.

RESULTS AND DISCUSSION

Assembly of five Stiff Stalk genomes

High-quality assemblies were generated for five Stiff Stalk founder lines B84, LH145, NKH8431, PHB47, and PHJ40 from approximately 124.2 million PacBio reads (Table 2). Assembly sizes ranging from 2.13 Gbp (NKH8431) to 2.18 Gbp (LH145) are comparable to previous PacBio assembly sizes of 2.13 Gbp, 2.2 Gbp, and 2.29 Gbp for B73 v4 (Jiao et al., 2017), Mo17 (Sun et al., 2018), and Ia453-*sh2* (Hu et al., 2021), respectively. Each Stiff Stalk assembly had N50 contig lengths ranging from 894 kbp (PHJ40) to 3.1 Mbp (B84) with the largest contig measuring 18.4 Mbp and N50 scaffold lengths exceeding 200 Mbp. On average, 94.5% of the assemblies were anchored to the ten maize chromosomes.

A high proportion of WGS reads aligned to their cognate assembly; greater than 99.8% of WGS reads aligned to the non-B73 Stiff Stalk genome assemblies, and 96.1% of B73 WGS reads aligned to the B73 v4 genome assembly (Supplementary Table S3.2). Properly paired reads accounted for 94.3% (B73) to 98.7% (PHJ40) of the total alignments. The proportion of reads mapping to multiple genomic locations ranged from 12.0% in the B73 library to 15.7% in the B84 libraries.

With respect to genic content, a high proportion of RNA-seq reads aligned to their cognate assembly, regardless of inbred or tissue (Supplementary Table S3.3). The average alignment rate of the RNA-seq reads from the five tissues to their cognate genome assemblies was greater than 93.0% for all inbreds. The B84 root tissue was the only library with low alignment rate (80.0% of reads aligned). A megablastn query of the B84 root tissue alignment

file against the NCBI nt nucleotide sequence database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/; accessed 11 Feb 2019) with an e-value threshold of 1×10^{-20}, 90% identity, and 50% coverage did not detect widespread contamination. Further investigation of the B84 root tissue alignment file revealed a large spike of deletions on the reverse read occurring in the 33rd sequencing cycle that may have negatively impacted alignment. All six Stiff Stalk genome assemblies (B73 v4, B84, LH145, NKH8431, PHB47, and PHJ40) showed a high proportion of complete BUSCO orthologs, with very few orthologs categorized as fragmented or missing (Supplementary Table S3.4). The Stiff Stalk assemblies contained comparable amounts of single-copy BUSCO orthologs, ranging from 1548 (95.9%) in B84 to 1558 (96.6%) in PHJ40. These metrics are comparable to other PacBio-derived maize assemblies for B73 v4 (Jiao et al., 2017), Mo17 (Sun et al., 2018), and Ia453-*sh2* (Hu et al., 2021), containing 1551, 1553, and 1562 single-copy orthologs, respectively. Furthermore, less than 3% of the ortholog set was classified as fragmented or missing in any Stiff Stalk genome assembly, reflecting a high coverage of genic space.

Transposable elements are one of the most difficult components to assemble in plant genomes due to their repetitiveness and low divergence (Ou et al., 2019). We evaluated the contiguity of the TE space using the LTR Assembly Index (LAI) software (Ou et al., 2018b). Relatively high LAI values were observed across the assemblies, with an average of 26.78 (Supplementary Table S3.5), which falls into the "gold" quality category, as previously defined (Ou et al., 2018b). Regional LAI values of the pseudomolecules were consistently high across each chromosome (Supplementary Figure S3.1). The LAI of the assembled chromosomes was, on average, 79 times higher than those of the unplaced scaffolds (Supplementary Figure S3.2),

indicating substantially decreased contiguity of the TE space in the unplaced scaffolds relative to those that were placed into chromosomes.

Transposable Element Composition

Transposable elements were annotated first based on structural features and then based on homology to a pan-Stiff Stalk TE library. The pan-Stiff Stalk TE library was constructed using the manually curated library from the MTEC (Schnable et al., 2009) as the base with the addition of novel TE sequences from each Stiff Stalk genome. In each of the assemblies, approximately 87% of the genome was annotated as TEs (Supplementary Table S3.6). LTR retrotransposons contributed the most (average of 75.69%), with *Gypsy* and *Copia* elements contributing 46.69% and 25.26% to the genome size, respectively (Supplementary Table S3.6). About 50,000 intact LTR retrotransposons were identified in each genome, and more than half of these (55.5%) were younger than 150,000 years old (Supplementary Figure S3.3), suggesting active amplification of LTR retrotransposons and a relatively short life cycle of these elements. DNA TEs contributed 11.12% to genome size on average, with *CACTA* and *Helitrons* representing the most sizable DNA TE superfamilies at 3.64% and 3.51% of genomic content, respectively (Supplementary Table S3.6). Non-TE interspersed repeats (i.e., centromere, subtelomere, rDNA, and knobs) contributed to only 0.23% of the assemblies, which is probably an underestimate due to challenges in assembling these repetitive sequences (Ou et al., 2020).

Annotation of six Stiff Stalk Genomes

The six Stiff Stalk genomes were annotated in parallel using *ab initio* gene predictions in combination with empirical, inbred-specific transcript evidence from a core set of diverse tissues

(leaf, internode, root, shoot, self-pollinated endosperm) (Table 3.3). This approach ensured that the resulting gene annotation for each Stiff Stalk inbred was not confounded by gene models and/or transcript evidence from other accessions which have been shown to differ significantly in maize (Hirsch et al., 2016) and *Arabidopsis thaliana* (Gan et al., 2011). In addition to the five ex-PVP inbreds described above, we also annotated the B73 v4 reference genome assembly (hereafter referred to as B73 v4 MSU). The current annotation of the B73 v4 assembly (Jiao et al., 2017) incorporates an enormous set of publicly available transcript sequences generated across multiple platforms from multiple inbreds using a MAKER-P pipeline that resulted in a significant over-annotation of gene model isoforms. For example, there are 161,680 working transcripts in the B73 v4 Gramene annotation (Jiao et al., 2017) yet 73,362 transcripts in the B73 v4 MSU annotation, a number comparable to the 72,635 to 75,124 transcripts present in the B84, LH145, NKH8431, PHB47, and PHJ40 genomes (Table 3.3). Therefore, any direct comparison of the B73 v4 Gramene annotation to the five Stiff Stalk genome described in this study would be confounded due to the nearly double the number of transcripts present in the B73 v4 Gramene annotation. Thus, through the use of a core set of representative tissues specific to each of the six Stiff Stalk genomes and a streamlined annotation pipeline, we have minimized the frequency of unsupported isoforms. Furthermore, this permits direct comparisons between all of the six Stiff Stalk genomes and a reduction of artifacts associated with differential annotation methods.

Genome Variation of six Stiff Stalk Genomes

Variation in the six Stiff Stalk assemblies was examined at the gene and genome level. First, the relationship between six Stiff Stalk inbreds, two inbreds outside the Stiff Stalk heterotic pool (Mo17, PH207), and *Sorghum bicolor* was determined using a cladogram generated from

orthologous groupings (Figure 3.1a). All branches had multiple sequence alignment support values of 100%. As expected, *S. bicolor* was distantly related to the maize lines, and Mo17 and PH207 inbreds clustered separately from the six Stiff Stalk inbreds. Among the Stiff Stalk inbreds, B73 and B84 were closely related, while the PHI inbreds PHB47 and PHJ40 clustered together separately from the other inbreds.

To better understand the Stiff Stalk pan-proteome, we examined the presence of orthologous and paralogous groups within the predicted proteomes of the six Stiff Stalks (Figure 3.1b). A total of 236,356 genes (97.90% of all input genes) were assigned to 37,866 orthologous and paralogous groups, while the remaining predicted proteins were considered singletons. Very few Stiff Stalk proteins were assigned to paralogous groups (0.47% to 1.01%) or classified as singletons (1.66% to 2.69%), further reflecting the similarities between their predicted proteomes. The 23,846 'core' orthologous groups containing at least one gene from all of the Stiff Stalks made up 55.54% of the orthologous and paralogous groups and singletons, while 31.83% of orthologous groups were missing orthologs from one or more Stiff Stalk and were considered 'shell' orthologous groups (Supplementary Table S3.7). The 'cloud' groups were composed of inbred-specific paralogous groups (n = 354) and singletons (n = 5,066) across all six inbreds (Supplementary Table S3.7). In terms of proteins, the 'core', 'shell', and 'cloud' orthologous groups contained 74.65%, 22.59%, and 2.76% of the total predicted proteins, respectively. Inbred line PHJ40 contained the most inbred-specific paralogous groups and proteins (n = 1,170 groups, 1,496 proteins), while B73 contained the fewest (n = 697 groups, 836 proteins).

Next, to look at the nucleotide sequence conservation among the Stiff Stalk genomes directly rather than protein level conservation, we aligned the coding sequence (CDS) of the high confidence representative gene models from each Stiff Stalk inbred to each Stiff Stalk genome assembly. Genes were considered present in an inbred if they aligned to a unique location or multiple locations in the target genome. As expected, cognate gene alignments showed the highest proportion of genes classified as present (average of 99.66%). Among the six Stiff Stalks, the lowest proportion of genes present occurred when aligning PHJ40 genes to B84 (89.21%) and the highest proportion of genes present occurred when aligning B73 v4 MSU genes to B84 (Supplementary Table S3.8). The PHJ40 and PHB47 gene sets contained slightly lower proportions of "present" genes (89.58% to 90.13%) when aligned to the other Stiff Stalk assemblies. Considering that the annotated genes in each of the six Stiff Stalks contained similar proportions of BUSCO-derived orthologs (Supplementary Table S3.4), the relatively low alignment of PHJ40 and PHB47 could reflect subtle divergence from the other Stiff Stalks. The Orthofinder cladogram supports this hypothesis, as PHB47 and PHJ40 were not in the same clade as B73, B84, LH145, and NKH8431 (Figure 3.1a). With respect to the Stiff Stalk pan-genome, of the 241,034 Stiff Stalk pan-genes that were present in at least one assembly, 80.38% were considered 'core' genes present in all six Stiff Stalks, and the 'shell' and 'cloud' proportions were 17.79% and 1.83%, respectively (Supplementary Table S3.9). The proportions of pan-gene designations are comparable to those from the Stiff Stalk pan-proteome analysis, yet a greater proportion of genes were classified as 'core' using the representative gene model CDS alignments compared to the orthologous pan-genes (80.38% and 74.65%, respectively) due to the inherent differences in nucleotide- and protein-level variation. The Stiff Stalk pan-genome analyses had substantially less cloud genes than reported previously in analyses of the pan-

genome of larger diversity panels (Hirsch et al., 2014; Gage et al., 2019) or in comparison of B73 to PH207 (Hirsch et al., 2016) consistent with the higher degree of diversity and divergence between those inbreds, respectively, relative to this panel composed solely of Stiff Stalks.

To better understand the relationship between the Stiff Stalk and other heterotic pool pan-genomes, we examined two additional inbred lines, PH207 and Mo17, which represent the Iodent and Lancaster heterotic pools, respectively. As the methods used to annotate Mo17, PH207 and the six stiff stalks differed, we limited our analyses of the pan-genome in the Stiff Stalk, Iodent, and non-Stiff Stalk heterotic pools to alignments of representative gene model coding sequences to the eight genome assemblies. At the gene level, the Stiff Stalk genes were less likely to be present in the PH207 and Mo17 assemblies and vice versa (Supplementary Table 3.8). Notably, only 78.70% of Stiff Stalk genes were present in the PH207 assembly compared to 87.03% of PH207 genes found among the Stiff Stalk assemblies, which may indicate true divergence of PH207 but also the incompleteness of the PH207 assembly which was generated from short reads (Hirsch et al., 2016). In comparison, 88.73% of Stiff Stalk genes aligned to the Mo17 assembly, with PHJ40 genes in particular aligning slightly more often to Mo17 (90.33%) than to the other Stiff Stalks (89.58%). Even so, 68.77% of genes were present in all eight inbred assemblies and considered ‘core,’ 13.84% were present in seven assemblies, and 4.73% were present in at least six assemblies; in total, 29.69% of the genes were present in two to seven assemblies representing ‘shell’ genes (Figure 3.1c; Supplementary Table S3.10). Overall, 98.46% of the genes were either ‘core’ or ‘shell’ in comparison to just 1.54% of the total transcripts which aligned to a single assembly (‘cloud’). Core genes, present in all eight assemblies, as well as shell genes present in seven assemblies, were longer on average than

genes found in six or fewer assemblies (Figure 3.1d), consistent with previous observations about gene length and membership in the pan-genome (Gordon et al., 2017). Differences in gene complement between heterotic pools have been hypothesized to contribute to the heterosis observed in hybrids yet incompleteness in the genome assemblies, especially in the case of PH207, and differences in gene annotation methods can impact precise detection of allelic variants resulting in over-estimations of the dispensable portion of the pan-genome. Future studies with a broader set of inbred lines from the non-Stiff Stalk and Iodent heterotic pools will permit assessment of the extent of inbred- and heterotic pool-specific genes.

For synteny analysis, B73 was selected as the reference Stiff Stalk genome to which the other Stiff Stalk assemblies were compared. As expected, B73 gene density was elevated on the arms of the chromosome with gene expression mirroring gene density (Figure 3.2). Collinear blocks were identified for each Stiff Stalk inbred compared to B73, revealing high levels of collinearity (Figure 3.2; Supplementary Table S3.11). In total, 1,178 (B84) to 1,737 (PHJ40) collinear blocks were detected across the five Stiff Stalks, containing 45,741 (PHJ40) to 53,895 (B84) syntenic gene pairs, or syntelogs. (Supplementary Table S3.11). The detection of approximately 500 more collinear blocks and 8,000 fewer syntelogs in inbred line PHJ40 is attributable to its distance from B73 and its more fragmented genome assembly. The collinear blocks composed of chromosome-chromosome alignments made up 61.95% (PHJ40) to 65.03% (B84) of the total collinear blocks in each Stiff Stalk and contained 77.05% (PHJ40) to 84.75% (B84) of all syntelogs, demonstrating the genic content of the Stiff Stalks is present on the assembled pseudomolecules rather than unplaced contigs. The mean and maximum number of genes in these collinear blocks was largely consistent, with four of the five assemblies averaging

56 syntelogs per block and a maximum block size of 4,376 syntelogs, compared to PHJ40 with an average of 33 syntelogs per block and a maximum block size of 1,120 syntelogs. The number of syntenic genes across B73 and each comparator Stiff Stalk detected by the synteny analysis ranged from 55,427 genes in the B73-PHJ40 comparison to 62,951 genes in B84-PHJ40 comparison. These genes made up 92.66% to 99.35% of all syntenic gene pairs found among chromosome-chromosome collinear blocks, which further reflects the high conservation of genic content among the Stiff Stalk inbred lines.

Structural variation between B73 and the five Stiff Stalk inbreds was primarily due to genomic deletions, insertions, inversions, and duplications with sizes ranging from small insertions of 31 bp up to inversions as large as 6.14 Mbp (Figure 3.3a). The total number of SVs detected ranged from 23,197 in B84 to 42,295 in PHJ40. The number of SVs categorized as deletions or insertions was influenced by relatedness to the B73 comparator; lines such as B84, LH145, and NKH8431 had fewer SVs relative to PHB47 and PHJ40, however, the proportion of SVs categorized as deletions was consistent across the five Stiff Stalks (69.13% to 75.08%) (Figure 3.3b). In a genomic context, deletions were the predominant SV across all five Stiff Stalks, representing 197.74 Mbp (9.28%) of the B84 assembly to 447.60 Mbp (20.78%) in PHJ40 which was the most fragmented assembly. We noted an enrichment of deletions in the 9 to 11 kbp size class (Figure 3.3a) and, upon inspection of TE annotations of deleted sequence, we found 60.6% of deletions in this range were fl-LTRs, which are typically 9 to 11 kbp in size. To test for fl-LTR enrichment, we extracted random sequences mimicking the exact length of these deletions in the B73 v4 genome and performed the same annotation with 10 iterations to find only 13.1% of random sequences were fl-LTRs. A Fisher's Exact test confirmed enrichment

of fl-LTRs in deletion SVs compared to random genomic sequence ($p < 0.00001$). Insertions represented 39 to 48 Mbp in four of the five Stiff Stalks, excepting PHJ40, which contained 97.65 Mbp of SVs categorized as insertions. Although few in number, inversions made up a substantial proportion of the Stiff Stalk nucleotide content (Figure 3.3c). Notably, LH145 contained 59.34 Mbp of inverted sequence (2.72% of the assembly), which was substantially greater than the other Stiff Stalks of which the next largest inversion content was 41.00 Mbp in line NKH8431 (1.93% of the assembly). The largest inverted region was found in NKH8431 (6.14 Mbp) on chromosome 4 at 96.76 Mbp. Duplicated SVs made up a small fraction of Stiff Stalk assemblies in terms of both number and nucleotide content.

Resistance Gene Diversity

Disease resistance genes are well documented as fast evolving gene families (Michelmore et al., 2013; Krattinger and Keller, 2016) and access to six Stiff Stalk genomes that arose through artificial selection provides a powerful dataset to understand the extent of diversity in a set of closely related genomes. The predicted proteomes of the six Stiff Stalks genomes were categorized into classes of resistance genes based on the detection of domains associated with disease resistance (Osuna-Cruz et al., 2018). The six Stiff Stalk inbreds had similar putative resistance gene profiles (Supplementary Table S3.12); in total, 19 unique classes of resistance genes were identified in the predicted proteomes from the six Stiff Stalk inbreds. The six Stiff Stalk predicted proteomes contained similar quantities of putative resistance genes, ranging from 1,818 in B73 to 1,903 in LH145 with kinases and receptor-like kinases representing approximately 49% and 27% of putative resistance genes, respectively. In comparison, proteins

classified as receptor-like kinases made up 42% and 36% of the putative resistance genes detected in Sorghum and Arabidopsis, respectively.

The 1,818 predicted B73 resistance genes were compared across the Stiff Stalks. As disease resistance genes can share significant sequence similarity, we used synteny to determine the presence of syntelogs between B73 and the five Stiff Stalk inbreds. Of the 1,818 putative B73 resistance genes, only 202 (11%) were unique to B73 (Supplementary Figure S3.4). When a B73 resistance gene was present in at least one of the five Stiff Stalks, the most common copy number was four instead of the expected five. Both biological and technical factors are likely contributing to this value, since PHJ40 is more distantly related to B73 compared to the other lines and also has a more fragmented assembly. Indeed, the number of resistance gene syntelogs to B73 ranged from 327 (PHJ40) to 1,485 (B84), representing 17 to 64% of the total syntelogs for their respective pairwise comparison. When PHJ40 was excluded from the analyses, the most common copy number was four, which corresponds to the number of non-B73 Stiff Stalks. Some B73 resistance genes were duplicated in the Stiff Stalk genomes, most notably a cluster of kinases near 188 Mbp on chromosome one (Supplementary Table S3.13); these B73 genes were annotated as wall-associated kinases and were highly expressed in the leaf tissue.

Presence-absence variation (PAV) has been well documented in maize (Springer et al., 2009; Lai et al., 2010; Hirsch et al., 2014, 2016). To highlight this phenomenon, we investigated a previously characterized gene conferring resistance to sugarcane mosaic virus, *ZmTrxh* (Liu et al., 2017). This gene is located on chromosome 6 near 24 Mbp in the B73 inbred line and is within a known PAV (Gustafson et al., 2018; Gage et al., 2019). *ZmTrxh* was present in a large

collinear block shared among B73 and three Stiff Stalk inbreds: B84, LH145, and PHB47 (Figure 3.4a). When the B73 ZmTrxh protein sequence was queried against the six Stiff Stalk genomes, no hits were detected in the NKH8431 and PHJ40 genome assemblies suggestive that it is a PAV in these two inbreds. Previous disease incidence scores indicate that SCMV resistance is quantitative, and that presence of *Scmv1* within the PAV is necessary but not sufficient for SCMV resistance (Gustafson et al., 2018). In contrast, a cluster of genes encoding the biosynthesis of DIMBOA (2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one) near 3.7 Mbp on chromosome four were completely conserved across all Stiff Stalk inbreds (Figure 3.4b). These findings further support the notion that general defense mechanisms such as DIMBOA biosynthesis conferring broad resistance across plant pathogens are more highly conserved compared to single-gene based disease resistance.

Founders and Conserved Regions in Descendants

We sought to determine the representation of the BSSS population within the five Stiff Stalk inbreds evaluated and a group of publicly released or commercial ex-PVP inbreds. B84 is directly from BSSS (HT)C7, and the four ex-PVP lines have one or more inbreds in their lineage derived directly from a version of BSSS. The founder inbreds are diverse amongst themselves, having only a few small regions that are shared by more than two lines, as exemplified by the founder haplotypes on chromosomes two and three (Figure 3.5a, remaining chromosomes in Supplementary Figures S3.5 and S3.6). Likewise, BSSSC0 inbreds show a mosaic of shared haplotypes with the founders on chromosomes two and three, and exhibit much shorter contiguous haplotypes, as expected after several generations of recombination and inbreeding (Figure 3.5b). Two founder lines are absent from our analysis, resulting in some BSSSC0 lines

containing haplotypes that are not present in the founder lines. For the publicly released and ex-PVP inbreds, haplotypes that are not found in the base BSSS population are plotted in white to facilitate visualization of BSSS haplotype conservation. The public inbreds have a greater diversity of haplotypes present compared to the ex-PVP inbreds, which exhibit a large reduction in diversity potentially due to the founder effects of commercial usage of B73 (Figure 3.5c, 3.5d). Haplotype blocks are largest, as measured in distance in base pairs, in pericentromeric regions, which is expected due to lower SNP density in the RNA-seq data and lower levels of recombination (Figure 3.5a). Several haplotypes move to fixation on both chromosome two and chromosome three, but only chromosome two shows significantly elevated F_{st} compared to the genome-wide average. Twenty four out of 109 blocks on chromosome two rank in the highest 10^{th} percentile of genome wide F_{st} values, having a value greater than 0.53, while only 1 out of 101 blocks on chromosome three ranks with high F_{st} (Figure 3.5e, Supplementary Table S3.14).

In Figure 3.5f, BSSS founder and BSSSC0 haplotypes are plotted for the five assembled Stiff Stalk genomes. As in the publicly released and ex-PVP lines, non BSSS haplotypes are plotted in white. As expected, B84 has high levels of conservation of the base BSSS population. Of the 900 total genome-wide blocks, 87.1% of blocks in B84 are from the founder or BSSSC0 lines. LH145 shares 64.7%, NKH8431 shares 57.6%, PHB47 shares 67.8%, and PHJ40 shares the least haplotype blocks with the base BSSS population at 29% (data not shown). There are several possible reasons for the haplotype blocks that are unique to B84 compared to the base BSSS population. In addition to the absence of two founder lines from our study, the unique haplotypes could be due to genotyping error, residual heterozygosity, mutation, or population contamination sometime during development or maintenance of the line. A range of 12.9%

(B84) to 71% (PHJ40) of the genome-wide haplotype blocks in the sequenced inbreds come from outside the base BSSS population, as demonstrated by the white segments in Figure 3.5f, which highlights the unique and diverse nature of these five lines despite their common placement in the Stiff Stalk heterotic pool.

Genome-wide identity by state (IBS) was calculated for each of the five lines with their respective closest Stiff Stalk founders. As expected, PHB47 has a high level of identity with its parent B37, where 73.1% of the 900 genome-wide haplotype windows have greater than 97% IBS. Despite this high level of IBS, identity is not distributed evenly in the genome, and seven of ten centromere containing regions are diverse between the two lines (Supplementary Figure S3.7). LH145 has high identity with its founder B14, which is found in the backgrounds of both of its parents, A632Ht and CM105. The pedigree of A632 (sans "Ht", Northern Corn Leaf Blight resistance) is B14 crossed to Mt42 with three backcrosses to B14, and B14 is also a direct parent to CM105 (npgsweb.ars-grin.gov). LH145 and B14 have high IBS in 64.4% of genome wide windows (Supplementary Figure S3.8). B84 shares 39.0% of IBS windows with B73 and has fewer and shorter conserved haplotypes compared to the direct relationship of PHB47 with B37 and LH145 with B14 (Supplementary Figure S3.9). NKH8431 has a higher level of IBS with B14 at 40.9% window sharing than B73 at 25.8%, which is expected due to its pedigree that includes two parents derived from B14 and one parent derived from B73 (Supplementary Figures S3.10 and S3.11). Finally, PHJ40 has IBS greater than 97% in 24.3% of genome windows with B37, with conserved haplotypes that are concentrated on chromosomes one, four, and nine (Supplementary Figure S3.12). The ancestral pedigrees of the proprietary inbreds used to generate PHJ40 are not known, but previous work indicates that B37 is a contributor to PHJ40,

with minor admixture from Lancaster and Oh43 germplasm (White et al., 2020). The lower level of IBS between B37 and PHJ40 is consistent with previous observations in this study that PHJ40 is more distantly related compared to the other Stiff Stalk inbreds, and agrees with our findings, as well (Figure 3.1a; Supplementary Tables S3.8 and S3.9).

As B73 is considered the reference genome for the maize community, we examined the relationship between SV and IBS regions in detail. Structural variants between B73 and B84 larger than 100,000 bp, including insertions, deletions, and inversions, were plotted for each chromosome (Supplementary Figure S3.9). Increased SV density was associated with decreased SNP IBS, as expected. Some regions with long stretches of high IBS do contain SVs, which could be due to the method of generating the SNPs by aligning RNA-Seq reads to the B73 reference, or decreased SNP density, such that the consecutive conserved SNPs fall on either side of the SV. Overall, SVs between B73 and B84 occur in non-conserved regions between the two lines.

Finally, we sought to determine the proportions of Stiff Stalk founders B14 and B37 that were present within the five Stiff Stalk inbreds that we sequenced. As previously noted, inbreds LH145, NKH8431, and B84 have direct relationships with B14, and 85.6% of the 900 genome wide windows have IBS greater than 0.97 between B14 and any of its related inbreds (Supplementary Figure S3.13). Similarly, 81.6% of the genome wide windows are conserved between B37 and its related inbreds PHB47, PHJ40, and B84 (Supplementary Figure S3.14). Thus, a high proportion of the genomic sequence of Stiff Stalk founders B14 and B37 is represented in the inbreds sequenced in this study.

CONCLUSION

Here we provide genomic resources for five historically important commercial Stiff Stalk inbred lines. High-quality *de novo* genome assemblies were generated with PacBio long read sequencing that contain near complete coverage of genic space as well as substantial repetitive content supporting the high-quality nature of the assemblies. Inbred-specific transcriptomes and gene annotations were independently generated using a core set of five tissues that permitted unconfounded comparisons of gene content across six key Stiff Stalk inbreds revealing broad similarity yet unique regions, reaffirming their usefulness in heterotic pattern breeding schemes.

The Stiff Stalk population has been an important source of seed parent germplasm for maize breeders in the public and private sectors since the mid 20th century. It is estimated that B14, B37 and B73 have an overall genetic contribution of 3.2%, 1.5%, and 11.7%, respectively, to inbred lines registered between 2004 and 2008 by the commercial breeding programs of Monsanto (now Bayer), Pioneer Hi-Bred, International (now Corteva) and Syngenta (Mikel, 2011). A study of ex-PVP inbreds estimated admixture of recently developed lines through kinship analysis, and found that of the 1,506 lines with kinship estimates, developed in the year 2000 or later, 15% had total Stiff Stalk admixture greater than 50%, and 33% of lines had Stiff Stalk admixture greater than 30% (White et al., 2020). Reciprocal recurrent selection in maize breeding has increased genetic distance between the Stiff Stalk and non-Stiff Stalk groups, as exemplified by increasing distance between the progressive cycles of BSSS and its partner population, the Iowa Corn Borer Synthetic No.1 (BSCBS) (Hinze et al., 2005). Complementation of deleterious, incompletely dominant alleles has been previously shown to drive hybrid vigor between heterotic groups (Yang et al., 2017a). Thus, selection for heterotic hybrids in the Stiff

Stalk by non-Stiff Stalk overall heterotic groups would be expected to drive divergent allele frequency between groups and reduce allelic variation within groups. Our results support that released inbreds, especially ex-PVP, contain quite limited allelic variation compared to that present in the original BSSS population, as represented by random BSSSC0 and founder inbreds in this study. Drift has previously been shown to play a major role in the population structure of the BSSS and the BSCBS (Gerke et al., 2015). Drift and founder effects likely contribute to the fixation of haplotypes that we observe, yet the fixation of rare haplotypes can contribute to genetic gain and phenotypic improvement if they contain favorable alleles for yield, heterosis, disease resistance, or agronomic improvement. As examples of changes observed through selection and drift, the combination of haplotypes spanning approximately 200 Mbp on chromosome two present in B73 did not exist in the base BSSS population, but reached fixation within a group of commercial germplasm, while a common haplotype present within the BSSS founders on chromosome three did not reach total fixation. Genetic diversity is vital to continued genetic improvement, and our results support that substantial genetic diversity remains within the broadly-defined Stiff Stalk heterotic pool. Empirical studies also indicate that yield heterosis may be found in non-canonical hybrids produced from inbreds from different Stiff Stalk sub-groups. In a diallel of thirteen inbreds from different heterotic patterns, hybrids PHB47 × NKH8431 and PHB47 × LH145 had the highest specific combining ability, suggesting that sufficient genetic diversity exists between the Stiff Stalk subgroups to form competitive hybrids, and certainly produce phenotypic segregation in crosses (White et al., 2020).

Founder haplotype conservation is demonstrated in each of the five Stiff Stalk inbreds assessed in this study. Selection on the BSSS population by Iowa State University followed by

incorporation into commercial breeding programs has led to the accumulation of alleles potentially important for yield and agronomic traits. These five Stiff Stalk inbreds represent founder alleles in elite contexts, which can aid the maize genetics community in the study of yield, quantitative traits, and adaptation to variable environments. In addition, the five Stiff Stalk inbreds span the genetic and institutional diversity of the pool, representing both heterotic subgroups and North American maize breeding entities in the 1980's, including Iowa State University, Pioneer Hi-Bred International, Holden's Foundation Seeds, and Northrup King. Thus, these lines can be used to study the population of alleles present within the Stiff Stalk heterotic group, which contribute to adaptation, genotype-by-environment interactions, and combining ability between the Stiff Stalk subgroups and non-Stiff Stalk subgroups. Substantial genetic and genomic diversity was identified within the assembled inbreds despite their highly selected and adapted nature, and diversity likely remains within the greater Stiff Stalk pool to be explored and utilized by maize breeders and geneticists.

ACKNOWLEDGEMENTS

This work was funded by the U.S. Department of Energy Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494) to CRB, SMK, NdL; the National Science Foundation Plant Genome Research Program IOS- 1546657 to CRB, IOS-1546727 to CNH; and IOS- 1546719 and IOS- 1822330 to MBH; and the National Institute of Food and Agriculture, United States Department of Agriculture Hatch 1013139 and 1022702 project to SMK. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No DE-AC02-05CH11231. We acknowledge the assistance of David Kudrna for contributions to the high molecular weight DNA isolation and the assistance of Brieanne Vaillancourt with sequence data management.

DATA AVAILABILITY

Raw genome sequence reads are available in the NCBI Sequence Read Archive under the BioProject identifiers listed in Supplementary Tables S3.1 and S3.2. The genome assemblies have been deposited in NCBI under accession numbers B84(JAGTWB000000000), LH145 (JAGTWC000000000), NKH8431 (JAGTWD000000000), PHB47 (JAGTWE000000000), and PHJ40 (JAGTWF000000000). RNA-seq reads used in this study were obtained from the NCBI Sequence Read Archive samples listed in Supplementary Table S3.3. The genome assemblies and associated annotation for all Stiff Stalk inbreds described in this study (B73, B84, LH145, NKH8431, PHB47, and PHJ40) are available from the Maize Genomics Resource at the University of Georgia (http://maize.uga.edu/) via a genome browser, BLAST search tool, and flat files. Large data files have been deposited at the Dryad Digital Repository (https://doi.org/10.5061/dryad.wh70rxwmw).

APPENDIX

FIGURES

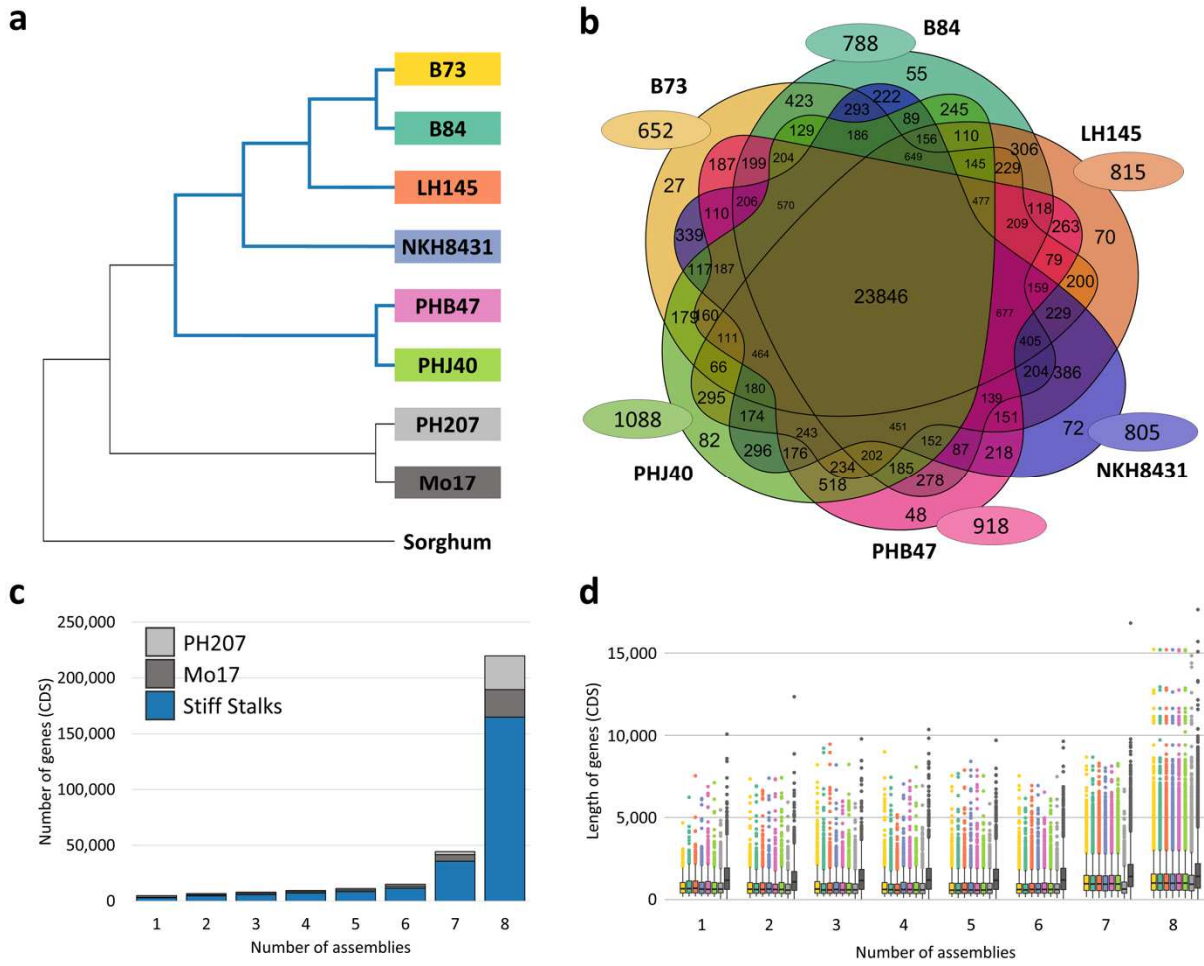


Figure 3.1. Stiff Stalk pan-proteome and pan-transcriptome.
Predicted proteomes for six Stiff Stalk inbreds, inbreds Mo17 and PH207, and *Sorghum bicolor* were assigned orthologous groups using Orthofinder v2.5.1 (Emms and Kelly, 2019). (a) Cladogram showing the relationships among proteomes. The cladogram was constructed and rooted from ancestral orthologous groups with the STAG and STRIDE algorithms (Emms and Kelly, 2018, 2017), respectively. Inbred lines belonging to the Stiff Stalk lineage are indicated in blue. All branches had multiple sequence alignment support values of 100%. (b) Venn diagram of orthologous and paralogous group occupancy across six Stiff Stalk inbreds. Intersections indicate orthologous groups containing at least one gene from a given Stiff Stalk. There were 23,846 'core' orthologous groups containing at least one protein from all Stiff Stalks, representing 55.57% of the total orthologous groups assigned (including singletons). Similarly, 31.83% of orthologous groups were missing at least one Stiff Stalk, and 12.62% of paralogous groups were unique to a Stiff Stalk inbred (i.e. inbred-specific paralogs plus singletons). The number of singletons for each Stiff Stalk inbred is shown in an ellipse overlaying the venn diagram. (c) High confidence representative coding sequences (CDS) of six Stiff Stalk inbreds,

Figure 3.1 (cont'd)

PH207, and Mo17 were aligned to the eight genome assemblies to assess presence/absence based on DNA sequence alignments with GMAP (Wu and Watanabe, 2005). (d) The length distribution of CDS considered present in one through eight assemblies are shown. Boxplots are colored according to inbred line as depicted in (a).

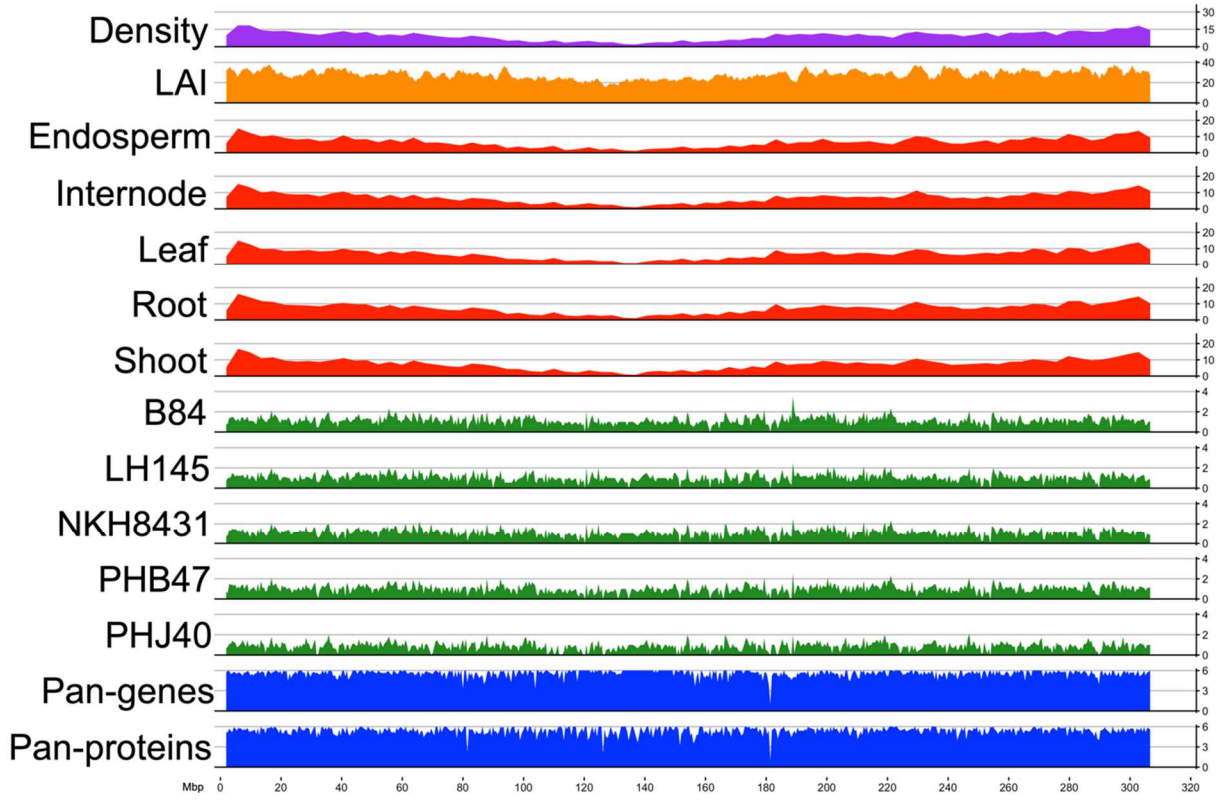


Figure 3.2. Gene density, gene expression, and syntelogs on *Zea mays* B73 chromosome one. The number of high confidence representative gene models is displayed in purple, LAI values are displayed in orange, and the number of expressed high confidence representative gene models across five tissues is displayed in red. The presence/absence of syntelogs among the other Stiff Stalks is displayed in green, and the average number of pan-genes and pan-proteins across the non-B73 Stiff Stalks are displayed in blue. Genic data were binned into 500 kbp non-overlapping windows for visualization.

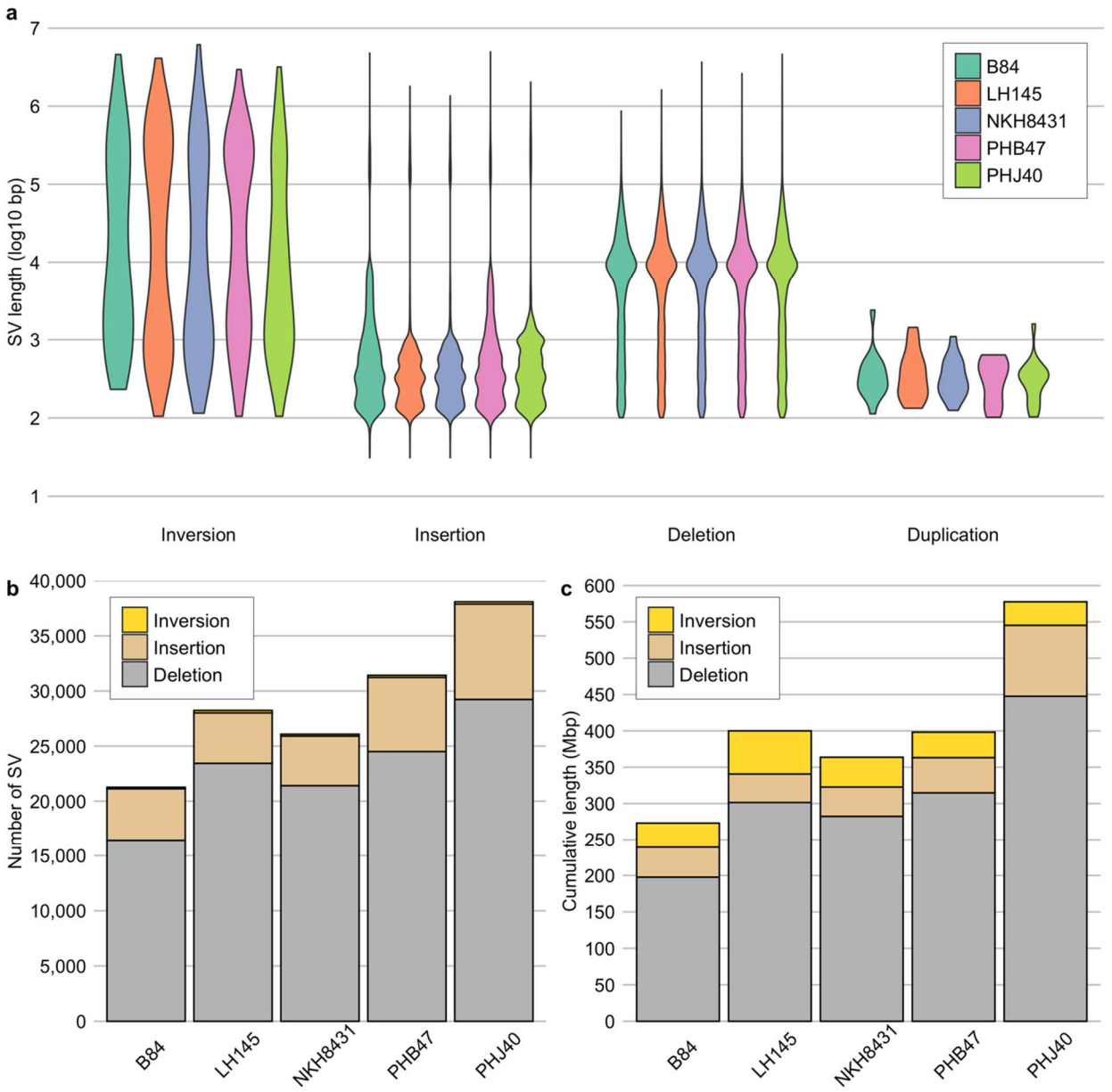


Figure 3.3. Structural variants across five Stiff Stalk assemblies.
(a) Distribution of $\log_{10}$-transformed lengths for the four most common structural variants detected. (b) Number of structural variants belonging to the four most common variant types. (c) Cumulative length of the four most common variant types.

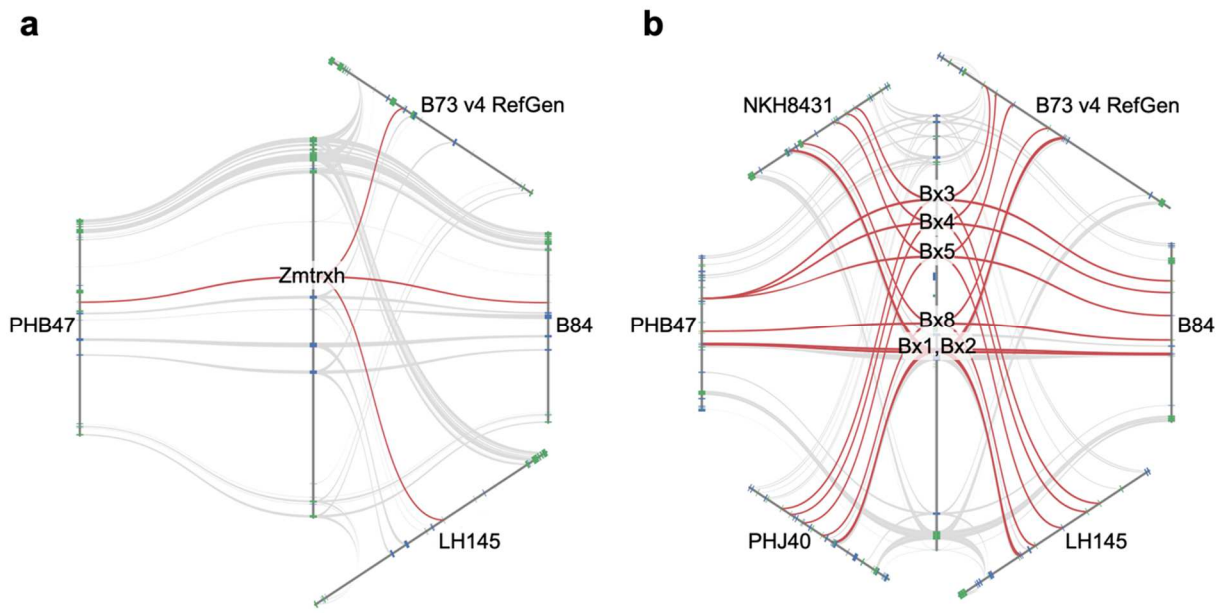


Figure 3.4. Resistance gene synteny among Stiff Stalks.
Coding sequences of the Stiff Stalk inbreds were aligned to the B73 v4 MSU annotation and syntenic regions were visualized with the python version of MCscan (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) implemented in the jcvi toolkit v1.1.7 with default parameters. (a) The Stiff Stalk inbreds exhibit presence-absence variation of the Zmtrxh locus near 24 Mbp on chromosome 6. (b) The Stiff Stalk inbreds exhibit complete conservation of DIMBOA gene cluster near 3.7 Mbp on chromosome 4. Relevant syntenic genes are highlighted by red connections, and adjacent syntenic genes are highlighted by grey connections. Genes on the forward and reverse strands are colored blue and green, respectively.

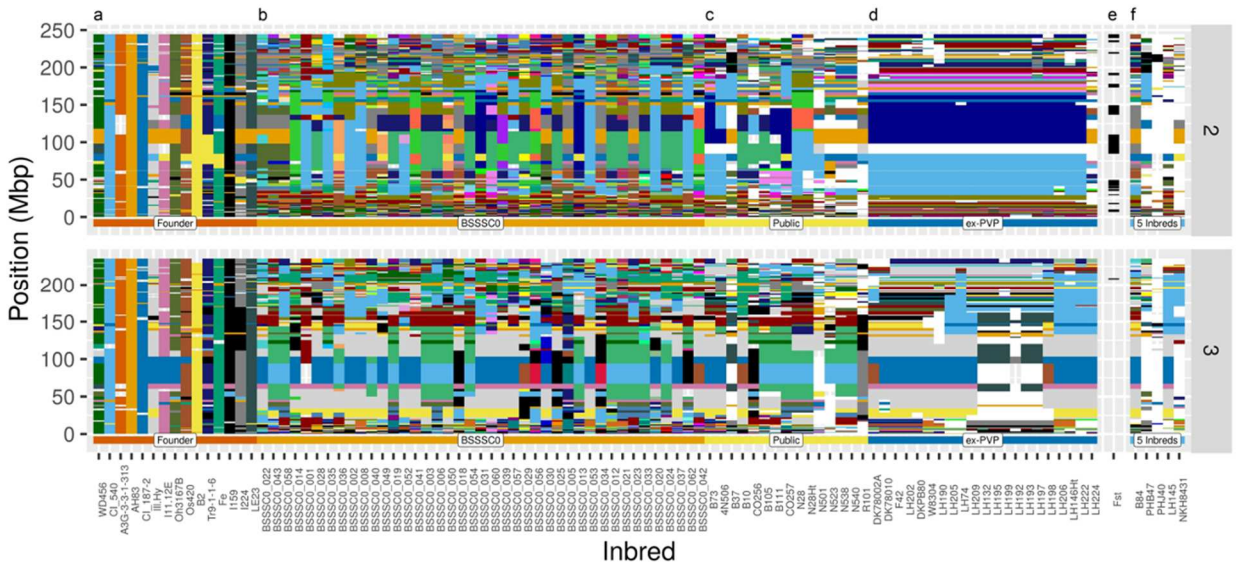


Figure 3.5. Stiff Stalk haplotypes and block F_{st} values.
(a,b) 960 SNP window haplotype blocks for founder inbreds (a) and unselected BSSSC0 inbreds (b) for chromosomes 2 and 3. (c,d) Conserved BSSS haplotypes for public releases from the BSSS populations (c) and highly related ex-PVP inbreds (d). Haplotypes not found in the founders or BSSSC0 lines are plotted in white. (e) Black boxes indicate haplotypes with binned maximum F_{st} values in the top 10^{th} percentile of genome-wide binned F_{st} values. F_{st} was calculated between the unselected lines, composed of the Founder and BSSSC0 inbreds, and selected lines, composed of the Public and ex-PVP inbreds. (f) Founder and BSSSC0 haplotypes present in the five assembled inbreds. Haplotypes with missing data are not plotted, showing the background of the plot. Major commercial inbred name prefixes: LH (Holden's Foundation Seeds, now owned by Bayer), DK (DeKalb Genetics Corporation, now owned by Bayer), PH (Pioneer Hi-Bred International, now owned by Corteva). For full descriptions of inbreds, see the Germplasm Resource Information Network database or (Mazaheri et al., 2019).

TABLES

Table 3.1. Origins of Stiff Stalk inbred lines described in this study.

Line	Originator	Place of Origin	Pedigree	PVP certificate or registration number	Date PVP or Registration Issued	PI number
B73	Iowa State University	Iowa, United States	Selected from advanced recurrent selection population (C5) of Iowa Stiff Stalk Synthetic (BSSS).	PL-17	01 Sep 1972	PI 550473
B84	Iowa State University	Iowa, United States	B84 is a selection from Iowa BSSS(HT)C7 [renamed BS13(S2)C0] that was tested as BS13(S2)CO-45-6-2-1-1.	PL-50	01 Jul 1979	PI 608767
LH145	Holden's Foundation Seed, Inc.	Iowa, United States	A632Ht x CM105	PVP 8300102	29 Jun 1984	PI 600959

Table 3.1 (cont'd)

Line	Originator	Place of Origin	Pedigree	PVP certificate or registration number	Date PVP or Registration Issued	PI number
NKH8431 (alias H8431)	Northrup, King & Company	Wisconsin, United States	(377 X B386) X 347 . All three parents are Northrup King proprietary lines originating from derivatives of Iowa Stiff Stalk Synthetic. Specifically, 377 derived from Iowa's B73, B386 derived from Minnesota's A632, and 347 derived from Iowa's B14	PVP 8800152	30 Nov 1988	PI 601610
PHB47 (alias B47)	Pioneer Hi-Bred International, Inc.	Minnesota, United States	B37 X SD105 specifically B37<3-XX#-SD105-#)F21323X11X. B37 is a public inbred line developed from Iowa Stiff Stalk Synthetic at Iowa State University. SD105 is an early public inbred line developed at South Dakota State University.	PVP 8300141	26 Oct 1984	PI 601009

Table 3.1 (cont'd)

Line	Originator	Place of Origin	Pedigree	PVP certificate or registration number	Date PVP or Registration Issued	PI number
PHJ40	Pioneer Hi-Bred International, Inc.	Ontario, Canada	B09 X B36 specifically B09/B36)X4122241X	PVP 8600133	31 Mar 1987	PI 601321

All information was obtained from the Germplasm Resource Information Network

Table 3.2. Genome assembly metrics for six Stiff Stalk inbreds.

Parameter	B73 v4.0[a]	B84 v1.0	LH145 v1.0	NKH8431 v1.0	PHB47 v1.0	PHJ40 v1.0
PacBio coverage	65x	88.4x	112.3x	113.7x	71.2x	85.4x
PacBio average read length (kbp)	11.7	8.4	6.1	6.2	6.9	6.1
PacBio reads (millions)	34.7	19.3	28.8	33.8	17.1	25.3
Scaffolds[b]						
GC content	46.9%	46.9%	46.9%	46.9%	46.9%	46.9%
Total number of scaffolds	265	291	1584	380	930	1547
Scaffold sequence (Mbp)	2,134.4	2,131.4	2,181.9	2,125.1	2,155.6	2,153.8
Scaffold N50 size (Mbp)	223.9	218.2	219.3	212.8	212.8	202.6
Scaffold L50 number	5	5	5	5	5	5
Breaks[c]	---	92	67	167	117	107
Joins[c]	---	1,184	2,115	1,626	2,908	2,702
Largest scaffold length (Mbp)	307	305	309	310	296	271
Contigs[b]						
Total number of contigs	2,785	1,475	3,699	2,006	3,841	4,250
Contig sequence (Mbp)	2,103.6	2,119.5	2,160.7	2,108.9	2,126.5	2,126.8
Contig N50 size (Mbp)	1.3	3.1	1.5	2.0	1.0	0.9
Contig L50 number	505	182	388	280	568	682
Largest contig length (Mbp)	7.3	18.4	10.0	15.6	8.5	7.8
Proportion of assembly on chromosomes[b]	98.7%	98.4%	95.0%	97.7%	95.1%	86.6%

[a]B73 v4 assembly was sourced from Gramene release 59

[b]Metrics do not include plastid sequences

[c]Misjoins identified by an abrupt change in B73 linkage group were corrected by creating breaks in the assembly

Table 3.3. Gene annotation metrics of six Stiff Stalk inbred genomes.

Stiff Stalk	B73 v4	B73 v4	B84[a]	LH145[a]	NKH8431[a]	PHB47[a]	PHJ40[a]
Annotation	Jiao et al. 2017	MSU[c]	MSU[c]	MSU[c]	MSU[c]	MSU[c]	MSU[c]
Working set							
Number of genes	49,085	49,986	50,861	52,133	50,732	50,982	51,335
Number of gene models	161,680	73,362	74,587	75,124	73,946	72,635	74,593
Mean gene model length (bp)	N/A	1,583	1,557	1,554	1,563	1,556	1,592
Median gene model length (bp)	N/A	1,350	1,321	1,319	1,327	1,309	1,335
High confidence set							
Number of genes	39,324	39,252	40,253	40,968	40,478	40,040	40,431
Number of gene models	131,319	62,091	63,430	63,451	63,114	61,156	63,110
Mean gene model length (bp)	N/A	1,766	1,730	1,736	1,734	1,742	1,777
Median gene model length (bp)	N/A	1,549	1,509	1,517	1,519	1,514	1,541
Representative/primary gene model							
Number of gene models[b]	39,498	39,252	40,253	40,968	40,478	40,040	40,431
Mean gene model length (bp)	1,584	1,476	1,455	1,457	1,452	1,464	1,470
Median gene model length (bp)	1,323	1,267	1,233	1,243	1,237	1,251	1,248

[a]Assembly provided in this paper

[b]B73 metrics include 174 plastid gene models

[c]Annotation provided in this paper

SUPPLEMENTARY MATERIALS

Supplementary Tables and Figures for Chapter 3 are included with the electronic version of the dissertation.

REFERENCES

REFERENCES

Beissinger, T.M., Hirsch, C.N., Vaillancourt, B., Deshpande, S., Barry, K., Buell, C.R., Kaeppler, S.M., Gianola, D., and de Leon, N. (2014). A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. Genetics **196**: 829–840.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23**: 2633–2635.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. **34**: 525–527.

Brohammer, A.B., Kono, T.J.Y., Springer, N.M., McGaugh, S.E., and Hirsch, C.N. (2018). The limited role of differential fractionation in genome content variation and function in maize (*Zea mays* L.) inbred lines. The Plant Journal **93**: 131–141.

Chang, C., Lu, J., Zhang, H.P., Ma, C.X., and Sun, G. (2015). Copy number variation of cytokinin oxidase gene Tackx4 associated with grain weight and chlorophyll content of flag leaf in common wheat. PLoS One **10**: 1–15.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W., and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods **10**: 563–569.

Clark, R.M., Tavaré, S., and Doebley, J. (2005). Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Mol. Biol. Evol. **22**: 2304–2312.

Coffman, S.M., Hufford, M.B., Andorf, C.M., and Lübberstedt, T. (2020). Haplotype structure in commercial maize breeding programs in relation to key founder lines. Theor. Appl. Genet. **133**: 547–561.

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, et al (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science **338**: 1206–1209

Doyle, J., J., Doyle, and L., J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin **19**: 11–15.

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

Emms, D.M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. **20**: 238.

Emms, D.M. and Kelly, S. (2018). STAG: Species Tree Inference from All Genes. bioRxiv.

Emms, D.M. and Kelly, S. (2017). STRIDE: Species Tree Root Inference from Gene Duplication Events. Mol. Biol. Evol. **34**: 3267–3278.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res **44**: D279-85

Gage, J.L., Vaillancourt, B., Hamilton, J.P., Manrique-Carpintero, N.C., Gustafson, T.J., Barry, K., Lipzen, A., Tracy, W.F., Mikel, M.A., Kaeppler, S.M., Buell, C.R., and de Leon, N. (2019). Multiple Maize Reference Genomes Impact the Identification of Variants by Genome-Wide Association Study in a Diverse Inbred Panel. The Plant Genome **12**: 180069.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, et al (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature **477**: 419–423

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, et al (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet **51**: 1044–1051

Gerke, J.P., Edwards, J.W., Guill, K.E., Ross-Ibarra, J., and McMullen, M.D. (2015). The Genomic Impacts of Drift and Selection for Hybrid Performance in Maize. Genetics **201**: 1201–1211.

Goodman, M.M. (1990). Genetic and Germ Plasm Stocks Worth Conserving. J. Hered. **81**: 11–16.

Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, et al (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. Nat Commun **8**: 2184

Graham, G.I., Wolff, D.W., and Stuber, C.W. (1997). Characterization of a Yield Quantitative Trait Locus on Chromosome Five of Maize by Fine Mapping. Crop Science **37**: 1601–1610.

Gustafson, T.J., de Leon, N., Kaeppler, S.M., and Tracy, W.F. (2018). Genetic Analysis of Sugarcane mosaic virus Resistance in the Wisconsin Diversity Panel of Maize. Crop Sci. **58**: 1853–1865.

Haas, B.J., Wortman, J.R., Ronning, C.M., Hannick, L.I., Smith, R.K., Jr, Maiti, R., Chan,

A.P., Yu, C., Farzad, M., Wu, D., White, O., and Town, C.D. (2005). Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. BMC Biol. **3**: 7.

Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, Seidel MA, Spannagl M, Marcon C, Ruban A, Urbany C, et al (2020) European maize genomes highlight intraspecies variation in repeat and gene content. Nat Genet **52**: 950–957

Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B, et al (2016) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated Solanum tuberosum. Plant Cell **28**: 388–405

Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B., Wiegert-Rininger, K., Wood, J.C., Douches, D.S., Farré, E.M., Veilleux, R.E., and Buell, C.R. (2017). Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proceedings of the National Academy of Sciences **114**: E9999–E10008.

Hinze, L.L., Kresovich, S., Nason, J.D., and Lamkey, K.R. (2005). Population Genetic Diversity in a Maize Reciprocal Recurrent Selection Program. Crop Sci. **45**: 2435–2442.

Hirsch CN, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, Barad O, Shem-Tov D, Baruch K, Lu F, Hernandez AG, et al (2016) Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell **28**: 2700–2714

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell **26**: 121–135

Hu Y, Colantonio V, Müller BSF, Leach KA, Nanni A, Finegan C, Wang B, Baseggio M, Newton CJ, Juhl EM, et al (2021) Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. Nat Commun **12**: 1227

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

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F.J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. **8**: 14061.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, et al (2017) Improved maize reference genome with single-molecule technologies. Nature **546**: 524–527

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

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

Krattinger, S.G. and Keller, B. (2016). Molecular genetics and evolution of disease resistance in cereals. New Phytol. **212**: 320–332.

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet **42**: 1027–1030

Larièpe A, Mangin B, Jasson S, Combes V, Dumas F, Jamin P, Lariagon C, Jolivot D, Madur D, Fiévet J, et al (2012) The genetic basis of heterosis: multiparental quantitative trait loci mapping reveals contrasted levels of apparent overdominance among traits of agronomical interest in maize (Zea mays L.). Genetics **190**: 795–811

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

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: 2078–2079.

Li Z, Zhou P, Della Coletta R, Zhang T, Brohammer AB, H O'Connor C, Vaillancourt B, Lipzen A, Daum C, Barry K, et al (2021) Single-parent expression drives dynamic gene expression complementation in maize hybrids. Plant J **105**: 93–107

Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, Gent JI, Llaca V, Woodhouse MR, Manchanda N, Presting GG, et al (2020) Gapless assembly of maize chromosomes using long-read technologies. Genome Biol **21**: 121

Liu, Q., Liu, H., Gong, Y., Tao, Y., Jiang, L., Zuo, W., Yang, Q., Ye, J., Lai, J., Wu, J., Lübberstedt, T., and Xu, M. (2017). An Atypical Thioredoxin Imparts Early Resistance to Sugarcane Mosaic Virus in Maize. Mol. Plant **10**: 483–497.

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, et al (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun **6**: 6914

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing

reads. EMBnet.journal **17**: 10–12.

Mazaheri M, Heckwolf M, Vaillancourt B, Gage JL, Burdo B, Heckwolf S, Barry K, Lipzen A, Ribeiro CB, Kono TJY, et al (2019) Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol **19**: 45

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**: 1297–1303.

Michelmore, R.W., Christopoulou, M., and Caldwell, K.S. (2013). Impacts of resistance gene genetics, function, and evolution on a durable future. Annu. Rev. Phytopathol. **51**: 291–319.

Mikel, M.A. (2011). Genetic composition of contemporary U.S. commercial dent corn germplasm. Crop Science **51**: 592–599.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. **41**: e121.

Osuna-Cruz, C.M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Aiese Cigliano, R., Sanseverino, W., and Ercolano, M.R. (2018). PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. Nucleic Acids Res. **46**: D1197–D1201.

Ou L, Li D, Lv J, Chen W, Zhang Z, Li X, Yang B, Zhou S, Yang S, Li W, et al (2018) Pan-genome of cultivated pepper (Capsicum) and its use in gene presence-absence variation analyses. New Phytol **220**: 360–363

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol **20**: 275

Ou S, Liu J, Chougule KM, Fungtammasan A, Seetharam AS, Stein JC, Llaca V, Manchanda N, Gilbert AM, Wei S, et al (2020) Effect of sequence depth and length in long-read assembly of the maize inbred NC358. Nat Commun **11**: 2288

Ou, S., Chen, J., and Jiang, N. (2018b). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. **46**: e126.

Ou, S. and Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. **176**: 1410–1422.

Pertea, G. and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Res. **9**: 304.

Pham, G.M., Hamilton, J.P., Wood, J.C., Burke, J.T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., and Buell, C.R. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. Gigascience **9**: giaa100.

Pucker, B., Holtgrawe, D., Stadermann, K.B., Frey, K., Huettel, B., Reinhardt, R., and Weisshaar, B. (2019). A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set. PLoS One **14**: e0216233.

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, et al (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet **45**: 1510–1515

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. **29**: 24–26.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326**: 1112–1115

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods **15**: 461–468.

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

Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, et al (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants **6**: 34–45

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet **5**: e1000734

Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, Barbazuk WB, Bass HW, Baruch K, Ben-Zvi G, et al (2018) The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet **50**: 1282–1288

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**: 637–644.

Sun S, Zhou Y, Chen J, Shi J, Zhao H, Zhao H, Song W, Zhang M, Cui Y, Dong X, et al (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet **50**: 1289–1295

Tracy, W.F. and Chandler, M.A. (2006). The Historical and Biological Basis of the Concept of Heterotic Patterns in Corn Belt Dent Maize. In Plant Breeding: The Arnel R. Hallauer International Symposium, K.R. Lamkey and M. Lee, eds (Blackwell Publishing), pp. 219–233.

Troyer, A.F. (1999). Background of U.S. Hybrid Corn. Crop Science **39**: 601–626.

[USC04] 7 USC Ch. 57: Plant Variety Protection (1970).

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature **557**: 43–49

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.-H., Jin, H., Marler, B., Guo, H., Kissinger, J.C., and Paterson, A.H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**: e49.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol. Biol. Evol. **35**: 543–548.

Weir, S., B., Cockerham, and . C.C. (1984). Estimating F-Statistics for the analysis of population structure. Evolution **38**: 1358–1370.

White, M.R., Mikel, M.A., de Leon, N., and Kaeppler, S.M. (2020). Diversity and heterotic patterns in North American proprietary dent maize germplasm. Crop Sci. **60**: 100–114.

Wu, T.D. and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21**: 1859–1875.

Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat. Methods **14**: 1072–1074.

Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E.S., Guill, K.E., McMullen, M.D., Mumm, R.H., and Ross-Ibarra, J. (2017a). Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLoS Genet. **13**: e1007019.

Yang N, Xu X-W, Wang R-R, Peng W-L, Cai L, Song J-M, Li W, Luo X, Niu L, Wang Y, et al (2017) Contributions of *Zea mays* subspecies mexicana haplotypes to modern maize. Nat Commun **8**: 1874

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

CHAPTER 4

GENETIC VARIATION IN A TEPARY BEAN (*PHASEOLUS ACUTIFOLIUS* L.) DIVERSITY PANEL REVEALS LOCI ASSOCIATED WITH AGRONOMIC TRAITS AND BIOTIC STRESS RESISTANCE

This chapter is in preparation for submission to The Plant Genome.

ABSTRACT

Tepary bean (*Phaseolus acutifolius* A. Gray), a relative of common bean (*Phaseolus vulgaris* L.), is indigenous to the arid climates of northern Mexico. In contrast to common bean, tepary bean is more well-adapted to higher temperatures that are increasingly likely due to climate change as well as exhibiting a wide range of resistance to biotic stressors. The tepary genome is highly syntenic to the common bean genome providing a foundation for discovery and breeding of agronomic traits between these two crop species. To date, a limited number of adaptive traits from tepary bean have been introgressed into common bean lines in spite of the hybridization barriers between these two species. To fully utilize tepary bean germplasm as a donor of adaptive traits, development of modern breeding resources and germplasm characterization is required. In this study, a diversity panel of 423 cultivated, weedy, and wild tepary bean accessions were genotyped and phenotyped revealing six subpopulations and differentiation of the subspecies of tepary bean. A genome-wide association study was performed with the 423-member diversity panel that illuminated loci and candidate genes underlying important agronomic traits which can be harnessed for not only tepary bean but also common bean improvement.

INTRODUCTION

Tepary bean (*Phaseolus acutifolius* A. Gray) is a dicotyledonous species in the Fabaceae (legume) family. Wild tepary bean has trifoliate leaves, a prostrate growth habit, and produces small, round seeds similar to those of its sister species, common or dry bean (*P. vulgaris* L.). There are two subspecies of *P. acutifolius,* subsp. *acutifolius* and subsp. *tenuifolius*. The subspecies can be distinguished by leaf morphology, with *tenuifolius* plants having smaller, narrower leaves compared to the larger, broader leaves of *acutifolius*, yet accurate classification of these two subspecies is challenging as the two subspecies can readily hybridize leading to a mixed genetic background and a range of leaf morphologies. Tepary bean is native to Mexico and is well-adapted to hot and dry conditions such as the Sonoran desert, and can be found across a range of elevations (International Center for Tropical Agriculture, 2022). In addition to abiotic stress resistance, tepary germplasm contains resistance to biotic stressors such as common bacterial blight (Drijfhout and Blok 1987; Singh and Mu oz 1999), Bean Common Mosaic Necrotic Virus (BCMNV; Porch, unpublished), and bean seed weevil (*Acanthoscelides obtectus*) (Mbogo et al., 2009; Kornegay and Cardona, 1991; Jiménez et al., 2017). As such, the resilience of tepary bean to biotic and abiotic stressors is a beneficial trait for plant breeders seeking germplasm and alleles that can improve to abiotic and biotic stress in these two *Phaseolus* spp.

Tepary Bean Breeding

Historical

Tepary beans were first domesticated by indigenous peoples inhabiting the arid region of what is now Mexico and the southwestern United States. These small-scale subsistence farmers grew tepary as part of a cropping system that included corn, sorghum, common bean, and squash

(Teiwes and Nabhan, 1983). According to traditional practices, tepary is planted into clusters or rows soon after the first or second summer rain, and plots are irrigated by floodwater runoff directed by ditches, berms, and brush. At harvest time, the vines are manually uprooted and dried for several days before being threshed, winnowed, dried again, then stored (Teiwes and Nabhan, 1983). Under this cropping system, traditional tepary producers likely selected for early maturity, pollen viability, drought tolerance, non-shattering pods, and larger seed size. These historical selections gave rise to two categories of cultivation status: semi-domesticated tepary plants growing in the wild ("weedy"), and cultivated landraces that can serve as a useful genetic reservoir for further tepary improvement.

Modern

Inter-specific hybridization is possible among *Phaseolus* species within gene pools of the the genus that are characterized based on the ease of hybridization (Harlan and Wet, 1971). The primary gene pool reflects within-species hybridization and is the most likely to produce viable progeny; however, F1 lethality can occur in some cases such as hybridization of *P. vulgaris* with parents from Andean and Middle American backgrounds (Singh and Ariel Gutiérrez, 1984; Gepts and Bliss, 1985; Hannah et al., 2007). In tepary bean, the secondary gene pool includes hybridization with the closely-related *P. parvifolius*, while the tertiary gene pool includes hybridization with the more widely cultivated common bean (*P. vulgaris*). Hybridization between tepary and common bean may require embryo rescue (Thomas and Waines, 1984), congruity backcrossing (Haghighi and Ascher, 1988), recurrent backcrossing (Mejía-Jiménez et al., 1994), or bridging lines (Barrera Lemus, 2021). Despite these challenges, traits from tepary have been successfully introgressed into common bean germplasm.

Common bean is the most consumed food legume globally (Broughton et al., 2003), but does not perform well under high heat or drought stress (Beebe et al., 2013; McClean et al., 2011; Porch et al., 2013a). This presents a challenge not only for current growers in tropical and subtropical areas where it is commonly consumed as a staple crop, but also future growers across the globe that will become hotter and/or drier due to climate change (Ramirez-Cabral et al., 2016). Tepary has also been used to introgress biotic stress resistance into common bean, such as common bacterial blight (*Xanthomonas* spp) (Costa and Rava, 2003; Scott and Michaels, 1992; Singh and Muñoz, 1999) and bruchid resistance (Myers and Kusolwa, 2011), although tepary lines with resistance to other pathogens including *Fusarium* wilt (Miklas et al., 1998), Bean Golden Yellow Mosaic Virus (BGYMV) (Miklas and Santiago, 1996), and rust (Pastor-Corrales et al., 2011) have been identified but have not yet been introgressed into common bean.

Tepary Bean Genomics

To date, tepary bean has been primarily used as a donor of a single or limited number of genes of interest to common bean breeders. However, efforts to breed and improve tepary itself have been made in recent years, aided by advances in *Phaseolus* genomics. The recent development of a tepary reference genome (Moghaddam et al., 2021) was preceded by similar efforts with commercially important relatives including soybean (Schmutz et al., 2010) and common bean (Schmutz et al., 2014; Vlasova et al., 2016). Genome assemblies and annotations for these the latter crops have been used for numerous breeding and research discoveries including quantitative trait mapping, genome architecture, genome-wide association (Moghaddam et al., 2016; Fang et al., 2017), evolution and domestication (Rendón-Anaya et al., 2017; Sedivy et al., 2017; Dong et al., 2021), and pan-genome studies (Liu et al., 2020), among

others. In addition, broader-scale studies in which diversity panels containing hundreds of individuals have been genotyped and phenotyped. For common bean, diversity panels representing the Andean (Cichy et al., 2015) and Middle American (Moghaddam et al., 2016) gene pools have been assembled and genotyped, facilitating numerous population genetics and trait mapping studies on agronomic traits (Kamfwa et al., 2015; Moghaddam et al., 2016), flooding tolerance (Soltani et al., 2018), anthracnose resistance (Zuiderveen et al., 2016), and cooking quality traits (Bassett et al., 2021; Katuuramu et al., 2018), among others.

Following the framework of common bean diversity panels, a tepary bean diversity panel (TDP, n=423) composed of cultivated, wild and weedy accession from both *P. acutifolius* subsp *acutifolius* and subsp. *tenuifolius* was assembled to facilitate research and breeding efforts. The panel was phenotyped for a suite of agronomic traits across a range of environments including heat stressed environments as well as examined for disease resistance to key pathogens. All accessions were genotyped permitting population genetics analyses and association mapping to enable an understanding of the diversity of the species and identification of loci associated with key agronomic traits that can be used to introgress into common bean.

MATERIALS AND METHODS

Tepary Diversity Panel Composition

The TDP consists of 423 tepary accessions and includes cultivated, wild, and weedy tepary lines collected from germplasm banks at the International Center for Tropical Agriculture (CIAT) and USDA National Plant Germplasm System (NGPS) (Table S4.1). Wild and weedy accessions were collected near the center of domestication in Mexico and surrounding areas, whereas cultivated accessions represent a combination of landraces and breeding lines developed at the USDA Tropical Agriculture Research Station (TARS) (Porch et al., 2022, 2013b).

Tepary Diversity Panel Growing Locations

The TDP was phenotyped for agronomic traits in trials that spanned a range of environmental conditions. Separate trials were planted in Juana Diaz, Puerto Rico in the summer and winter growing seasons to collect agronomic data (seed weight, seed size, yield) under heat and drought stress, respectively. Seed harvested from the heat trial was phenotyped for cooking quality traits. Root rot trials were undertaken in Isabella, Puerto Rico in a field site with endemic root rot pressure and low fertility. An additional trial was conducted in Fort Collins, Colorado at a location with *Fusarium solani* root rot pressure. Lastly, a trial was conducted in Tegucigalpa, Honduras to collect disease resistance ratings and leaf morphology measurements.

Tepary Diversity Panel Phenotyping

Seed weight was measured as the weight of 100 randomly-selected seeds from each growing location and also as areas and perimeters of seeds grown in the Juana Diaz drought trial calculated by the image processing software SmartGrain. Seedcoat color was measured from

seeds grown in the Isabela root rot trial using SmartGrain to obtain CIELAB values *L**, *a**, and *b** that correspond to the perceived lightness, redness-greenness, and blueness-yellowness, respectively (International Commission on Illumination, 2008). Disease and pest measurements were visually assigned according to interval rating scales.

DNA Isolation and Quantification

Genomic DNA for genotyping-by-sequencing (GBS) library construction was isolated by harvesting ~50 mg of leaf tissue (~1cm^2) from actively expanding trifoliate leaves, lyophilizing for 48 h, grinding to a fine powder using acid-washed silica beads (OPS Diagnostics) and a TissueLyser II (Qiagen), and then by using a DNeasy 96 Plant Kit (Qiagen) according to the manufacturer's instructions. The total volume for DNA elution was 40 µL. Double-stranded DNA was quantified with the QuantiFluor dsDNA Dye System (Promega) and a Quantus Fluorometer (Promega) according to the manufacturer's instructions. The DNA from each sample was diluted to 5 ng µL^{-1} with nuclease free water and arrayed in PCR plates in preparation for GBS library construction.

Genotyping-by-Sequencing Library Construction and Sequencing

Three separate Genotyping-by-Sequencing (GBS) libraries were constructed to ensure that all available germplasm was sequenced sufficiently. The DNA samples for each of the entries were assigned an *Ape*KI GBS barcode adapter according to the key file (Table S4.2). The GBS libraries were prepared using the methylation-sensitive restriction enzyme *Ape*KI and were constructed using previously published protocols (Elshire et al., 2011) as optimized for use with *P. vulgaris* in which 1.5 ng of each adapter was used per 50 ng of sample DNA (Hart and

Griffiths, 2015). The libraries were quantified, validated on an Agilent 2100 Bioanalyzer (Agilent Technologies), and sequenced at the Weill Cornell Medical College Genomics Resources Core Facility. The first library (TDP plates 1-4, 384-plex) was sequenced on four lanes of an Illumina HiSeq 2500 to obtain single-end 101 nt reads. The second library (TDP plates 5 & 6, 152-plex) and the third library (TDP plates 7&8, 152-plex) were sequenced on four lanes each of an Illumina NextSeq 500 to obtain single-end 75 nt reads.

Genotyping-by-Sequencing Data Processing

Raw GBS reads were quality inspected using FastQC v0.11.9 (Andrews, 2010). GBS reads were processed with the TASSEL GBSv2 pipeline v5.2.44 (Glaubitz et al., 2014) using default parameters unless otherwise stated. Reads with a minimum base quality score of 20 (-mnQS 20; default: 0) were aligned to the masked tepary reference genome pseudomolecules (Pacu.CVR.asm.hm.fa; (Moghaddam et al., 2021) using BWA-MEM v0.7.17 (Li, 2013) and filtered to include only those with a minimum MAPQ score of 20 (-minMAPQ 20; default: 0). A total of 207,154 variants were called from the aligned tags, and quality metrics were used to further filter the dataset. Specifically, variants with a F_{IT} less than 0.8 were discarded; 131,425 final variants were retained.

Variants were then filtered sequentially by converting sites to biallelic, removing insertion-deletion variants, removing SNPs with a call rate below 50%, and removing SNPs with a minor allele frequency (MAF) less than 1% using VCFtools v0.1.16 (Danecek et al., 2011) and BCFtools v1.13 (Danecek et al., 2021). SNPs with sample depths greater than the 99th percentile were set to missing on a per-individual basis. Heterozygous calls were set to missing prior to

imputation with BEAGLE v5.2 (Browning et al., 2018) using default parameters with the exception that the effective population size was decreased from 1,000,000 to 10,000 to accommodate the autogamous nature of *P. acutifolius*. A subset of accessions was sequenced more than once to obtain sufficient coverage (n=15); for these accessions, only the replicate with the higher average depth was retained after confirming congruence of the technical sequencing replicates. Additionally, accessions determined to be non-tepary (n=36) were removed by examining their placement in a neighbor-joining tree and Principal Component Analysis (PCA) plots. After removing these accessions from the VCF file, SNPs were removed if they had a MAF less than 1%. The final imputed VCF contained 423 tepary accessions with 53,877 SNPs with no missing calls.

Population Structure

A Q matrix representing population structure was calculated for the SNP dataset containing 423 tepary accessions and 53,877 SNPs using fastStructure v1.0 (Raj et al., 2014) with K=6 subpopulations, Admixture proportions for the accessions were visualized using the python script distruct.py bundled with the fastStructure program.

Genome-Wide Association Study

A genome-wide association study (GWAS) was conducted using MLM, MLMM (Segura et al., 2012), FarmCPU (Liu et al., 2016), and BLINK (Huang et al., 2019) methods implemented in the GAPIT R package v3 (Wang and Zhang, 2021). Population structure was controlled using six principal components, and kinship was controlled using a kinship matrix generated by the

methods of (VanRaden, 2008). The threshold for statistical significance was set at $\alpha = 1.9e\text{-}06$ based on the number of effective markers determined by SimpleM (Gao et al., 2008).

RESULTS AND DISCUSSION

Tepary Genetic Diversity

The TDP contains a diverse set of accessions across a range of seedcoat coloring, cultivation history, and geographical origin (Figure 4.1). Tepary accessions were genotyped by aligning their sequenced reads to the genome assembly of Frijol Bayo, a white-seeded tepary cultivar (Moghaddam et al., 2021). Called variants were thoroughly filtered before imputation to obtain a fully-imputed VCF file containing high confidence genome-wide SNPs. SNPs were enriched in euchromatic regions due to the ApeKI restriction enzyme used in GBS library preparation (Figure 4.2). Several non-tepary accessions were also sequenced as part of this study (Table S4.3); these outgroup accessions were used to identify questionable tepary accessions having genetic similarity to the outgroup species as determined by PCA (Figure 4.3). In total, 32 known outgroup accessions and 4 aberrant tepary accessions were excluded, resulting in a final VCF file that contained 423 tepary accessions and 53,877 SNPs.

Principal component analysis of the TDP dataset revealed clusters based on cultivation status and subspecies (Figures 4.3 and 4.4). PC1 explained 25.7% of the genetic variation and distinguished cultivated, weedy, and wild accessions. Subspecies differentiation by PCA was more subtle and was likely confounded due to the hybridization between wild subspecies and the absence of the *tenuifolius* subspecies in cultivated germplasm. In addition to PCA, the kinship matrix clearly delineated cultivated and wild accessions (Figure 4.5), in agreement with previous literature describing cultivated tepary as having a narrow genetic base (Mwale et al., 2020).

Subpopulations

Subpopulation structure was determined using K=6 subpopulations. These subpopulations roughly corresponded to geographic origin and cultivation status (Figure 4.1). Subpopulation 1 was composed of wild tepary collected in the Durango region of Mexico, subpopulation 3 included wild *tenuifolius* subspecies from southeastern Arizona, and subpopulation 4 categorized tepary from Central American countries. Tepary collected between the Gulf of California and the Sierra Madre Occidental mountain range belonged to subpopulations 5 (cultivated *acutifolius* subspecies) and 6 (wild and weedy *acutifolius* subspecies). Subpopulation 2 was a mixture of wild tepary from both *acutifolius* and *tenuifolius* subspecies.

Genome-Wide Association Study

Seed Size

One of the most striking differences between tepary and common bean is the difference in seed size. The weight of a 100 seed subsample, referred to as 100 seed weight is a widely-used measurement of seed size in common bean breeding, with beans in small-sized market classes such as navy and black beans typically having minimum weights around 18g per 100 seed (Evan Wright, pers. comm.). For comparison, the distribution of 100 seed weights in the TDP ranged from 0.5g (TDP-302) to 23.4g (TDP-297) (Figure 4.6), emphasizing considerable potential for improvement of this key trait in current and future tepary bean breeding efforts. Several QTL for 100 seed weight were identified across multiple algorithms and chromosomes. The most significant SNP, S03_10737408, was located on chromosome 5 at approximately 10.74 Mbp,

while other significant QTL for 100 seed weight were found on chromosomes 1, 2, 5, 6, and 8 (Table S4.4).

Seed area and perimeter are additional measurements of seed size. These traits were phenotyped from seed harvested from the Juana Diaz drought trial. QTL for seed area and perimeter colocalized to chromosome 2 at 25.34 Mbp (S02_25348260) and chromosome 8 at 47.43 Mbp (S08_47432373) (Figure 4.7). These QTL were detected by BLINK and FarmCPU algorithms and were supported by somewhat weaker significance scores from other algorithms like MLM and MLMM. Interestingly, the seed area and perimeter QTL on chromosome 8 was located near a SNP (S08_47156593) associated with 100 seed weight from the Juana Diaz heat trial. Additionally, a cluster of three SNPs with physical positions ranging from 1.43 to 5.59 Mbp on chromosome 8 were significantly associated with seed area.

Maturity

Phenological processes including inflorescence and maturity are important for tepary breeders seeking to introduce tepary lines that can be grown at higher latitudes with longer day lengths. No significant QTL were detected for days to flowering, however, the most significant QTL for days to maturity was detected by all algorithms (FarmCPU, BLINK, MLM, MLMM) on chromosome 8 near 11.36 Mbp for plants grown in Juana Diaz under drought stress (Figure 4.8). Under the MLM model, the associated SNP, S08_11364947, had a minor allele frequency of 1.4% with the minor allele (T) hastening maturity by 0.2 days. This QTL should be interpreted with caution since the plants were grown under substantial heat stress and the phenotypic extremes for maturity were within half of a day. The SNP S01_46538999 was also associated for

days to maturity, but only in the Honduras location. This QTL is near the previously-identified tepary homolog of the *Terminal Flowering 1* (*TFL1*) gene, Phacu.CVR.001G234100, located at 49.92 Mbp in the tepary reference genome (Moghaddam et al., 2021). Other QTL associated with days to maturity were detected on chromosomes 2, 4, and 7 (Figure 4.8; Table S4.5).

Seedcoat Color

Common beans grown as dry edible beans are categorized into distinct market classes based on seed characteristics like size, shape, and seedcoat coloring. For tepary beans to be similarly commercialized, it would be useful to have a better understanding of the loci governing seedcoat coloring. The most significant QTL for seedcoat color was detected on chromosome 6 near 17.2 Mbp across mapping algorithms for *L** and *b** values. Although *L** and *b** measure distinct aspects of the perceived color spectra, they were previously determined to be highly correlated for seedcoat color of cooked beans (Bornowski et al., 2020b); thus, the colocalization of *L** and *b** QTL is unsurprising. The SNP with the lowest p-value for this QTL (S06_17220748) had a minor allele frequency of 9.2% and was located among other significant SNPs with effect sizes of 15.8-16.2 and 6.9-7.0 on *L** and *b**, respectively. Other seedcoat color QTL with support across multiple algorithms can be found in Table S4.6. In common bean, the master regulator of seed coat color, the *P* gene, is located on chromosome 7 (McClean et al., 2018). A chi-square test of S07_31144955 genotype calls and seed coat color was significant ($p = 0.000128$) (Table S4.7). To investigate if the tepary *P* gene was detected in the tepary GWAS, BLASTP was used to identify the tepary *P* gene homologs, Phacu.CVR.007G206800 and Phacu.CVR.007G206900, located near 32.3 Mbp on chromosome 7, two gene models that should be merged to form the complete *P* gene. This physical location of the tepary *P* gene

homologs was nearby a FarmCPU SNP (S07_31144955) at 31.1 Mbp associated with *b** (Figure 4.9).

Biotic Stress Resistance

Tepary has a history of utilization as a donor species, whereby disease and pest resistance traits are introgressed into common bean breeding lines. In the TDP, QTL were found for resistance to bruchid damage during storage, virus resistance, and common bacterial blight (strains 484A and 3353). The most significant SNP bestowing bruchid resistance over a 60-day interval was S07_32760033 and was associated with a 19.5% reduction in the percentage of damaged seed (Table S4.8). A subset of the TDP was screened for resistance to Bean Common Mosaic Virus (BCMV) strain NL3-D using an enzyme-linked immunosorbent assay (ELISA), and three QTL for this trait were concordant across algorithms. The most significant SNP (S01_51567184) was detected on chromosome 1 near 51.56 Mbp and was associated with a 40% reduction in likelihood of a positive ELISA test. The other QTL for BCMV resistance were found on chromosomes 11 and 10, providing a 19% and 27% reduction, respectively (Table S4.9). A prior, preliminary study on BCMV resistance among the wild TDP accessions also described QTL on chromosomes 1 and 11 (Ana Vargas, unpublished), though it is unclear if the QTL are identical with those found in the present study. Lastly, QTL for resistance to common bacterial blight (CBB) were found on chromosomes 7 and 8. SNPs S07_14152143 and S07_34038504 were the most significant and were detected across CBB strains and mapping algorithms (Table S4.10). The beneficial alleles for these two SNPs lowered CBB ratings by 1.8 to 2.1 on a 1-9 rating scale.

CONCLUSION

Tepary bean is a climate resilient crop with substantial resistances to biotic and abiotic stressors. Although tepary landraces have been continuously selected by native people over the years, modern breeding techniques are necessary for this crop to quickly reach its potential as a commercial food legume. This study utilized a diversity panel of tepary beans to uncover genomic regions associated with important agronomic traits that can be targeted by breeders. The majority of these traits were found to be under polygenic control, and comparisons with homologous genes and regions in common bean and soybean lend further support to our results. The genotyping resources generated by this study serve as a basis for assessing tepary diversity in future genetic studies and provide candidate molecular markers for tepary breeding. The effects of climate change are likely to be variable, and additional investment in resilient crops such as tepary may hold the answer to future food security.

ACKNOWLEDGMENTS

N.B. was supported by a fellowship from the National Science Foundation (NSF) Research Traineeship - Integrated training Model in Plant And Compu-Tational Sciences (IMPACTs) Program. Acknowledgments go out to Dr. Timothy Porch and Dr. John Hart for their efforts in assembling, sequencing, and phenotyping the tepary diversity panel.

APPENDIX

FIGURES

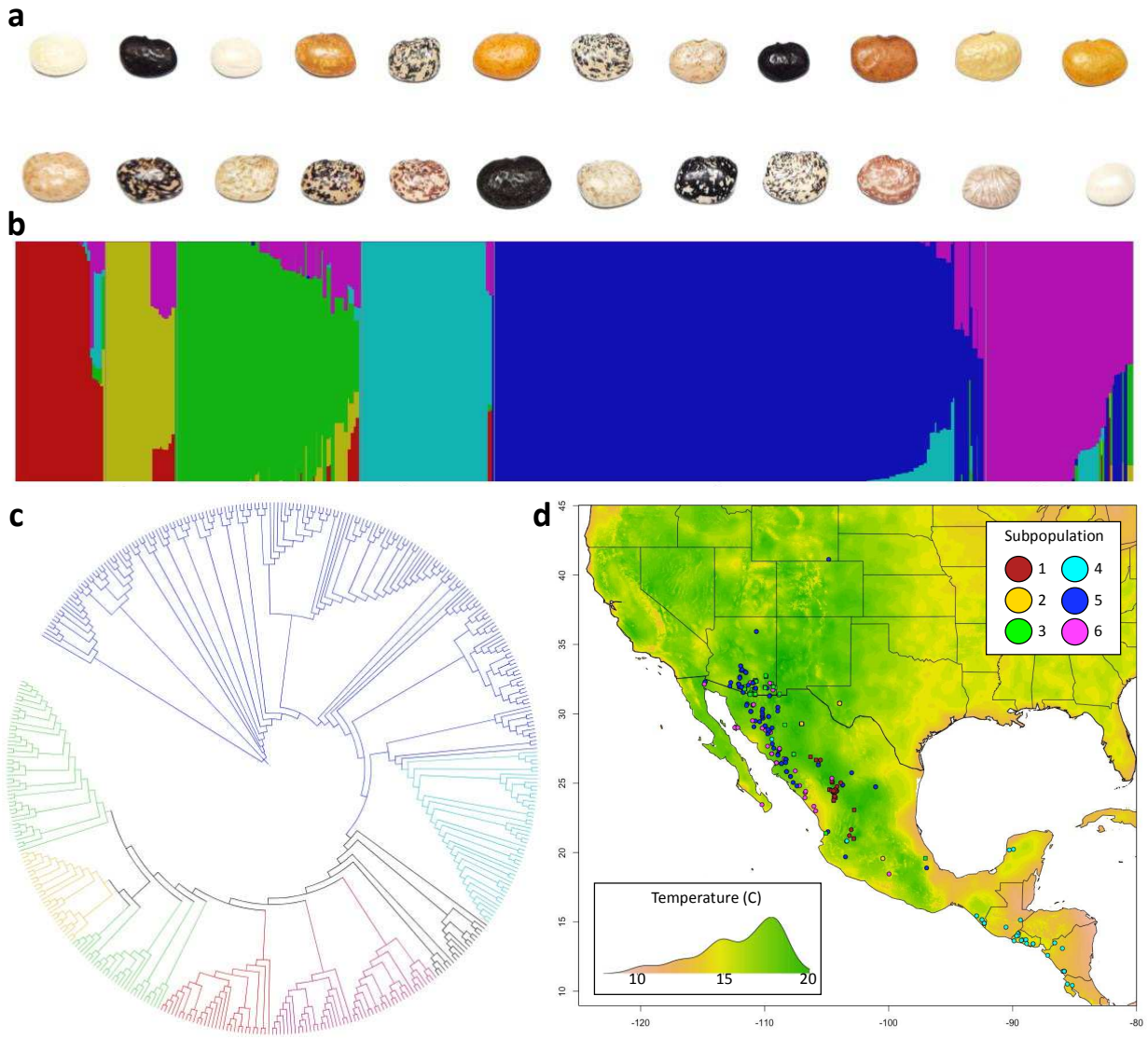


Figure 4.1. Phenotypic and genetic diversity in the Tepary Diversity Panel.

a) Tepary bean seed coat diversity.
b) Subpopulation structure and admixture of the Tepary Diversity Panel. Structure was determined using K=6 subpopulations with the software fastStructure v1.0 (Raj et al., 2014). Subpopulations were influenced by geographic origin and cultivation status. Population identities are provided in panel d.
c) Genetic distance tree of Tepary Diversity Panel accessions. Population identities are provided in panel d. Branch lengths are not to scale.
d) Geographic origin of Tepary Diversity Panel accessions. Collection locations were overlaid on a map of monthly diurnal temperatures (Celsius). Subpopulation coloration corresponds to coloring in panels b and c. Most accessions were collected in Mexico and surrounding areas. Accessions with Canadian, African, or unknown origin are not shown.

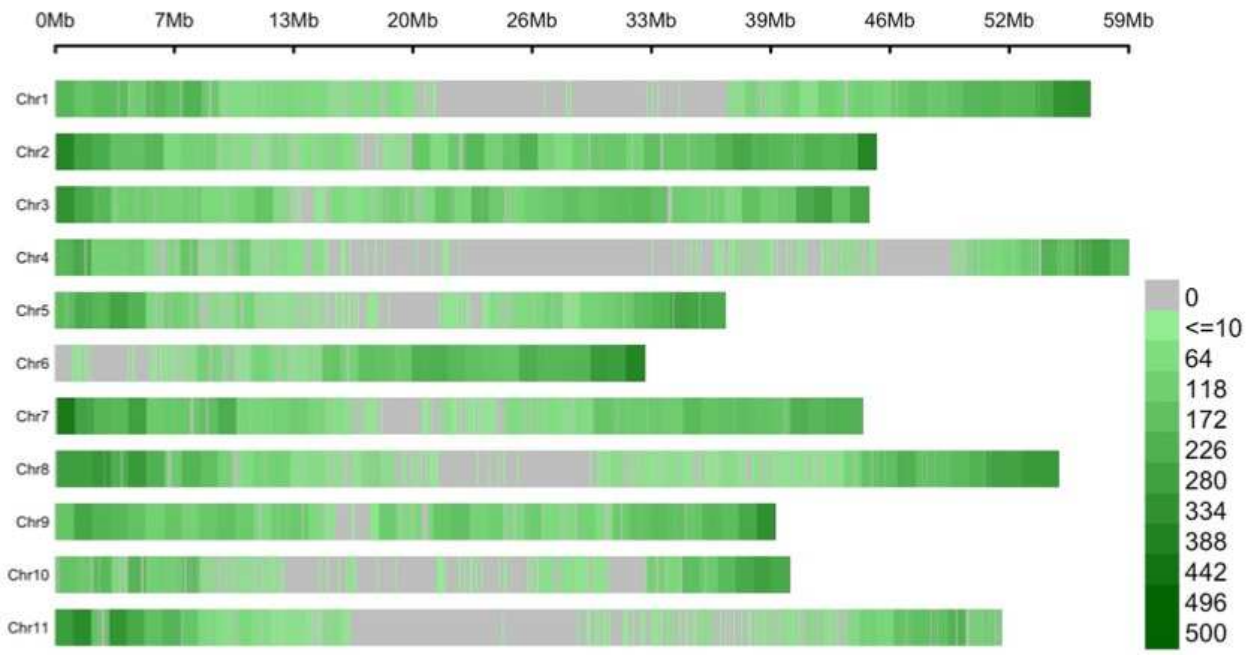


Figure 4.2. Genome-wide distribution of SNP markers in the Tepary Diversity Panel. All SNPs (n=53,877) were binned according to non-overlapping bins of 1 Mbp. SNPs were enriched in euchromatic regions due to the use of the methylation-sensitive *Ape*KI enzyme in genotyping-by-sequencing library preparation.

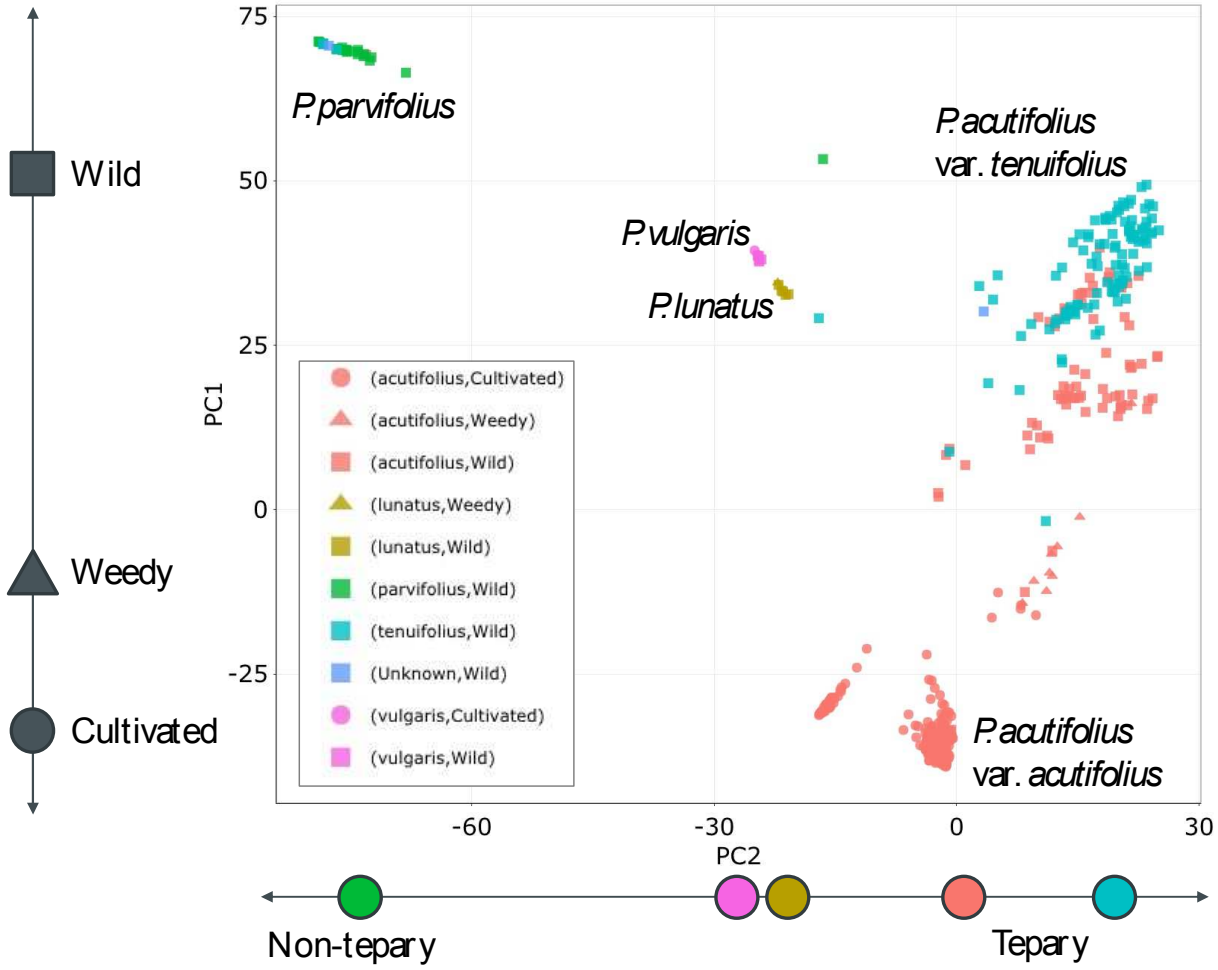


Figure 4.3. Principal component plot of the Tepary Diversity Panel and outgroup accessions.
Principal component 1 (y-axis) distinguished cultivation status according to wild, weedy, and cultivated accessions. Principal component 2 (x-axis) distinguished *Phaseolus* species.

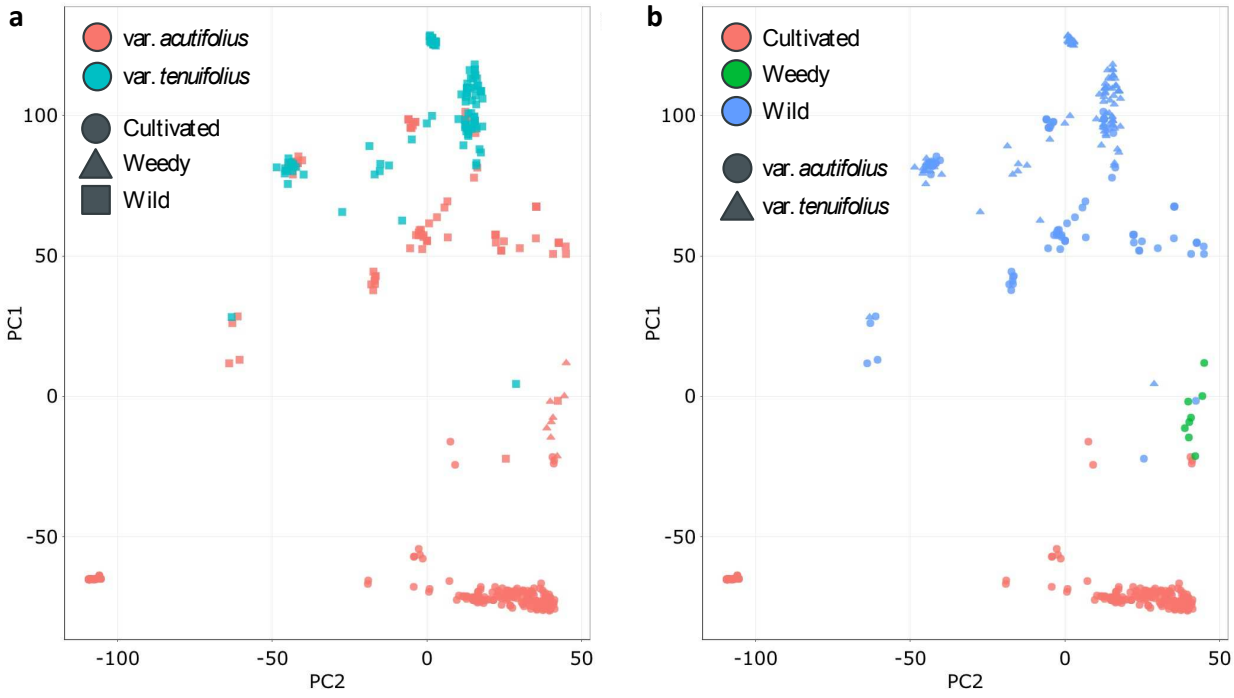


Figure 4.4. Principal component plot of the Tepary Diversity Panel.
Principal component 1 explained 25.7% of the total genetic variation and distinguished between cultivated and wild tepary accessions.

a) PCs 1 and 2 colored according to *P. acutifolius* subspecies (*acutifolius* or *tenuifolius*).
b) PCs 1 and 2 colored according to cultivation status (cultivated, weedy, or wild).

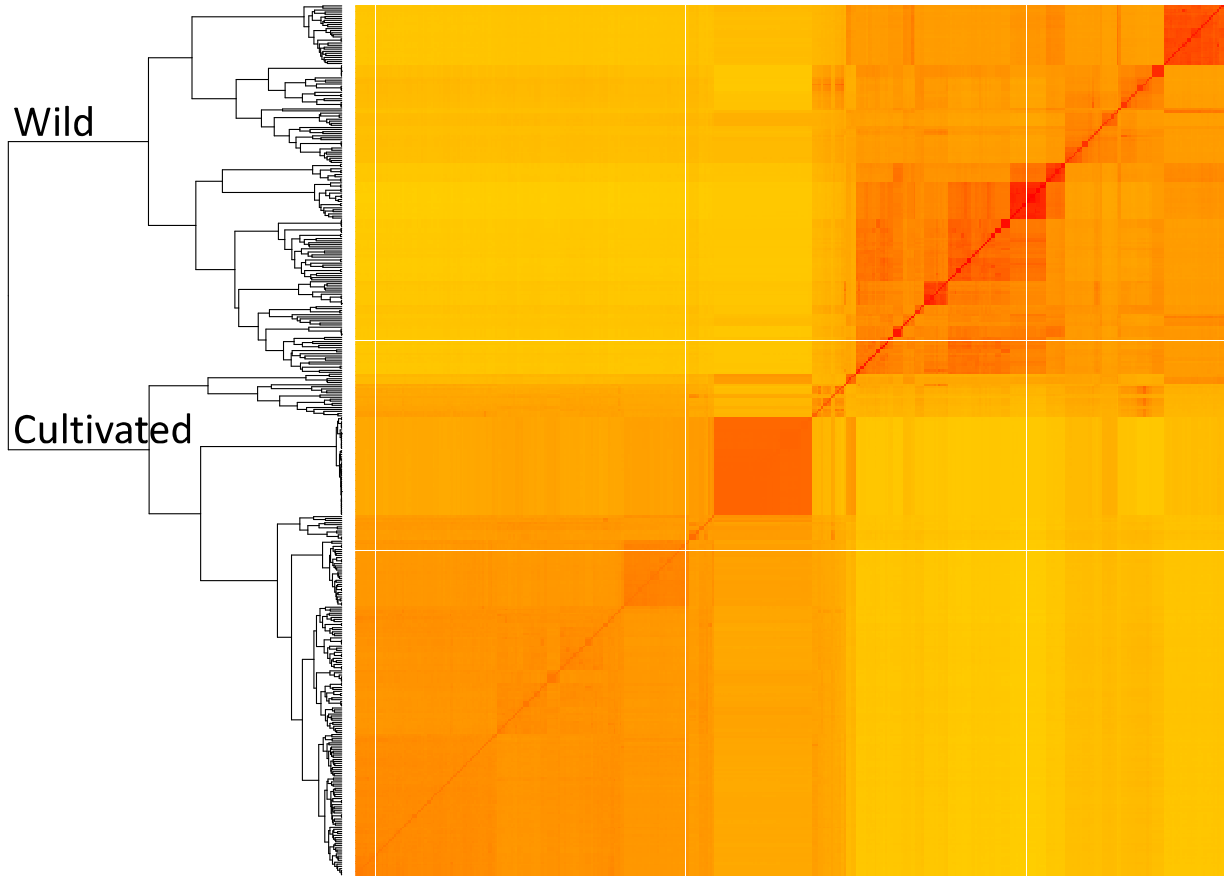


Figure 4.5. Tepary Diversity Panel kinship matrix heatmap.
Kinship was calculated among the 423 tepary accessions using the methods of VanRaden (2008), revealing the narrow genetic base of current cultivated tepary accessions.

Figure 4.6. Distribution of 100 seed weights among the Tepary Diversity Panel accessions. 100 seed weight is the weight of 100 randomly-selected seeds in grams. Values shown are averages across all growing locations. Generally, cultivated accessions have larger seeds.

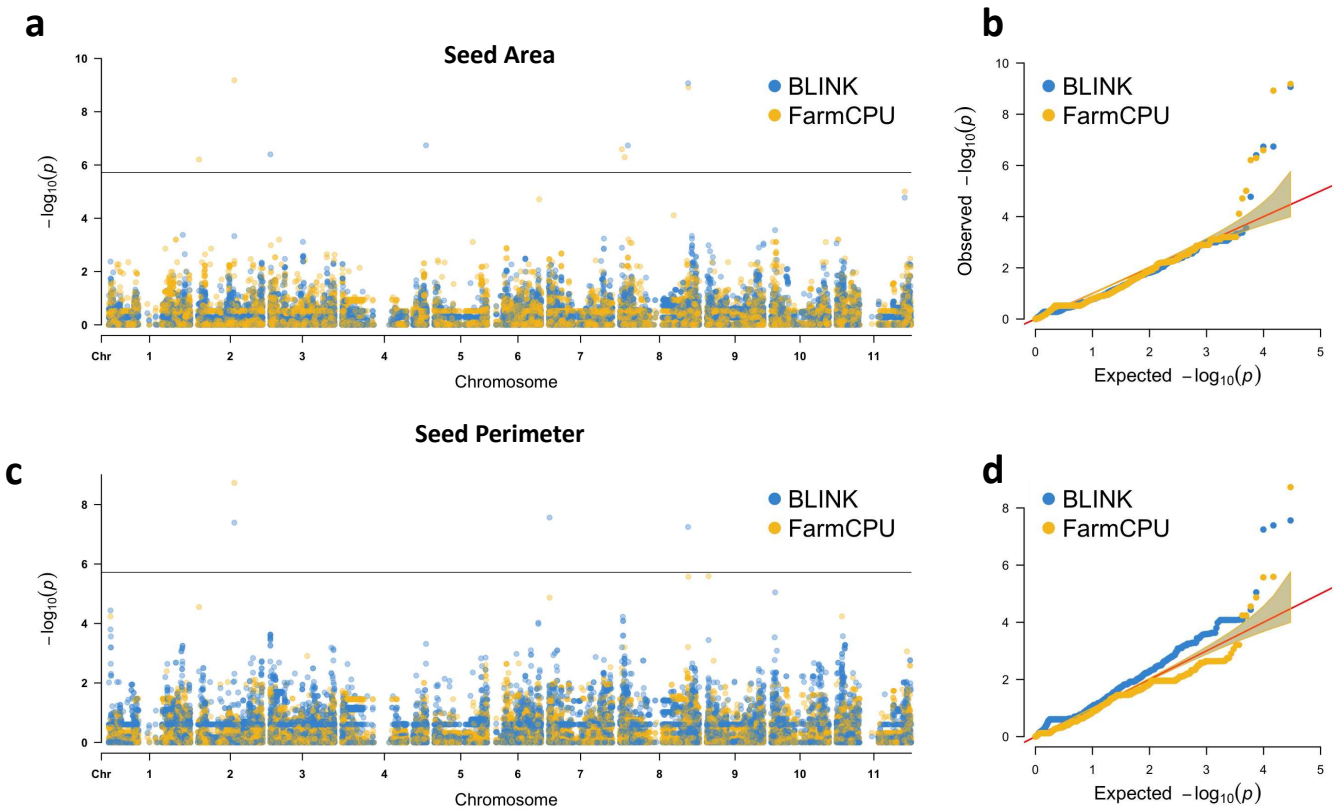
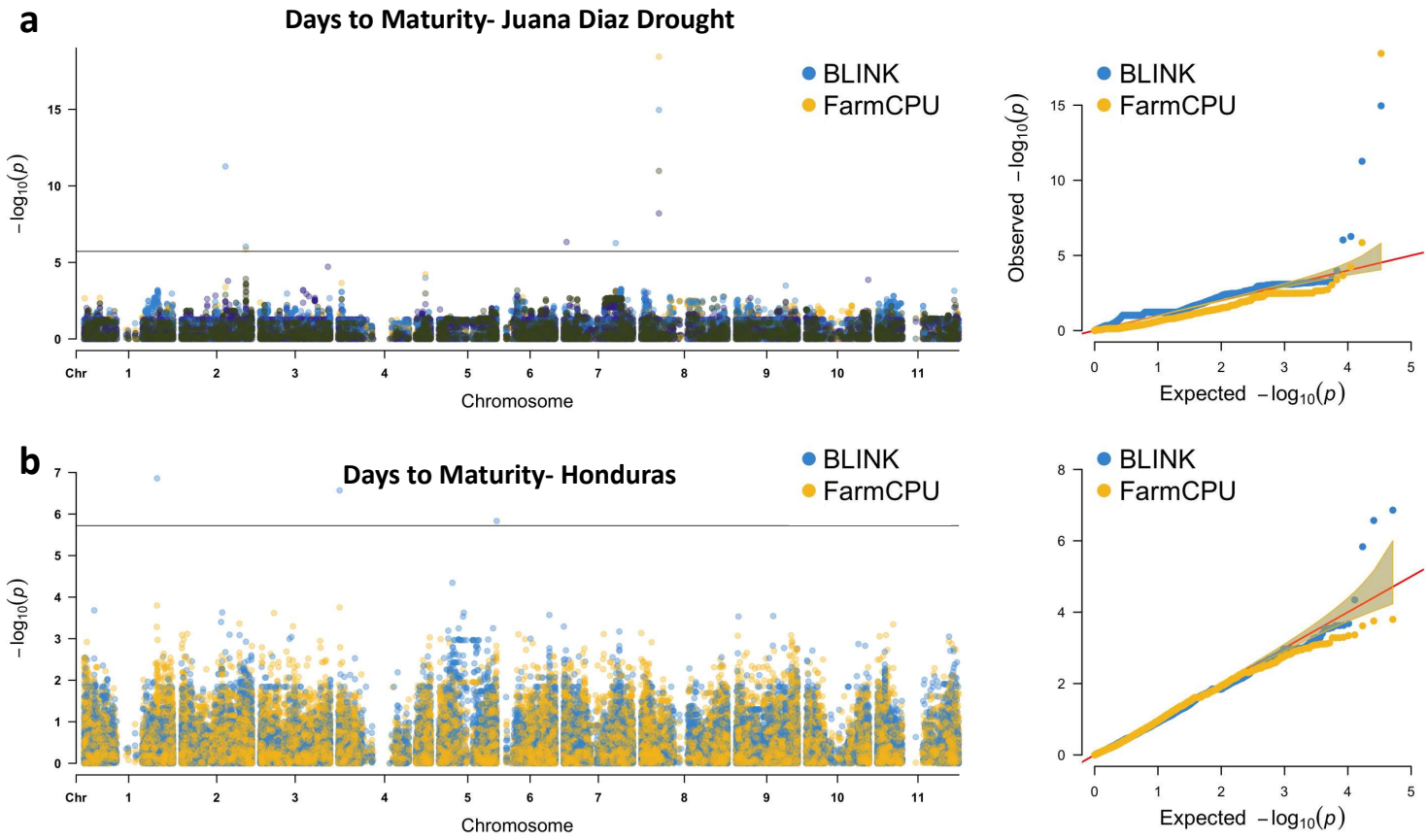


Figure 4.7. Manhattan plots for seed area and perimeter.
GWA was conducted in the GAPIT R package v3 (Wang and Zhang, 2021) using two algorithms. The significance threshold is based on the number of effective markers estimated by SimpleM. (Gao et al. 2008). SNPs connected by the vertical dashed line were detected across multiple algorithms and seed size traits.

a) Manhattan plot for seed area.
b) Quantile-Quantile plot for seed area.
c) Manhattan plot for seed perimeter.
d) Quantile-Quantile plot for seed perimeter.



e)

Figure 4.8. Manhattan plots for tepary maturity traits.
SNPs are colored according to mapping algorithm used. The significance threshold is based on the number of effective markers estimated by SimpleM. (Gao et al. 2008).

a) Days to maturity as observed at Juana Diaz under drought stress.
b) Days to maturity as observed at the Honduras growing location.

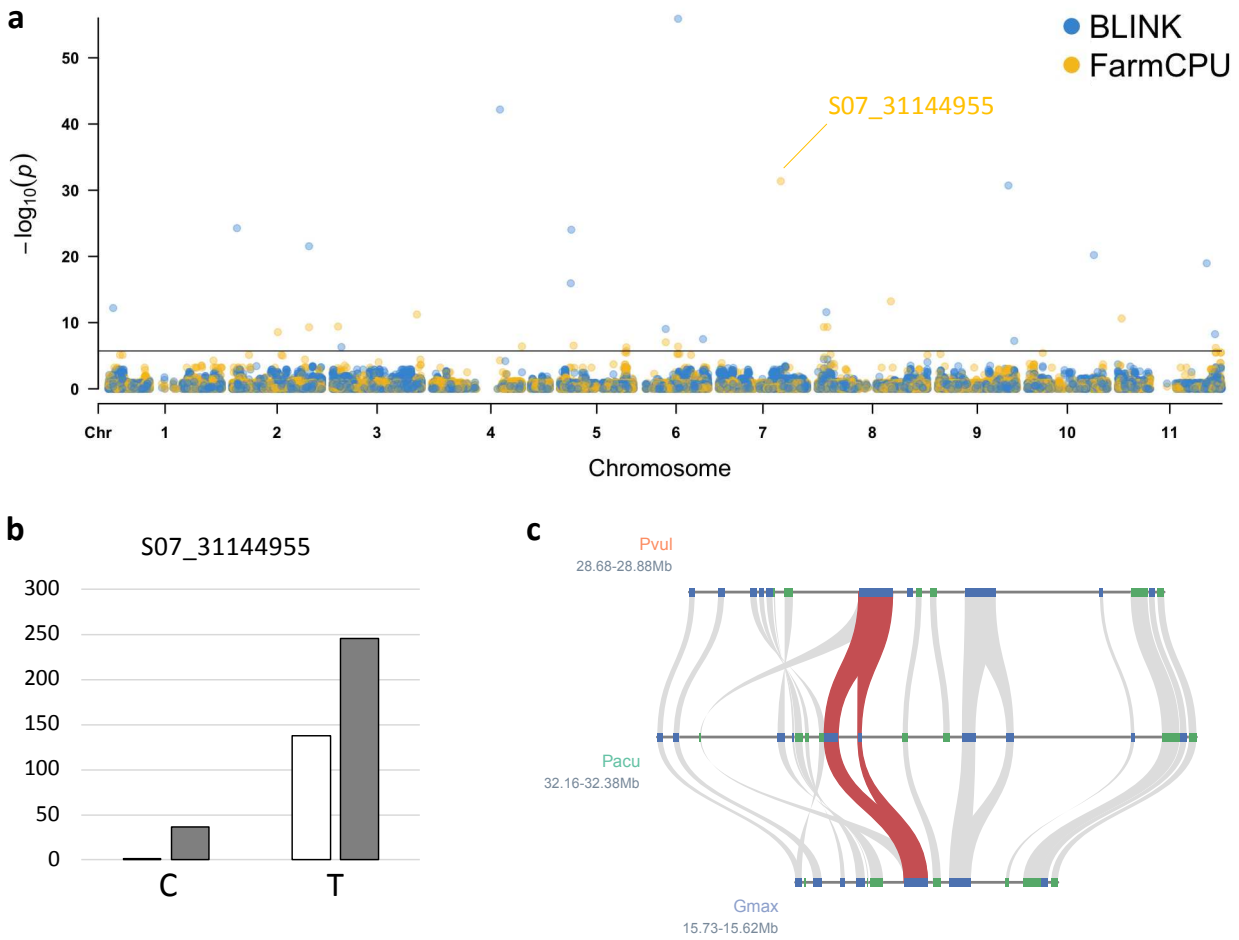


Figure 4.9. Genome-wide association of seed coat color.

a) Manhattan plot for seed coat color *b** value. The *b** value corresponds to the perception of blueness or yellowness. SNP S07_31144955 is located in the vicinity of the tepary *P* gene homologs, Phacu.CVR.007G206800.1 and Phacu.CVR.007G206900.1.

b) The genotype calls for SNP S07_31144955 are significantly associated with the presence or absence of seed coat color in the TDP (Chi-square test; p = 0.000128). The number of tepary accessions with white and colored seed coats is shown by the white and grey bars, respectively.

c) The tepary *P* homologs (Pacu) are within a conserved syntenic block with common bean (Pvul) and soybean (Gmax). Genes are depicted as blue or green boxes and syntenic relationships are depicted with grey connecting lines. The syntenic relationships of the *P* gene homologs are depicted with red connecting lines.

SUPPLEMENTARY MATERIALS

Supplementary Tables for Chapter 4 are included with the electronic version of the dissertation.

REFERENCES

REFERENCES

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Barrera Lemus, S. (2021). The University of Nebraska - Lincoln ProQuest Dissertations Publishing **28865047**.

Bassett, A., Kamfwa, K., Ambachew, D., and Cichy, K. (2021). Genetic variability and genome-wide association analysis of flavor and texture in cooked beans (*Phaseolus vulgaris* L.). Züchter Genet. Breed. Res. **134**: 959–978.

Beebe, S.E., Rao, I.M., Blair, M.W., and Acosta-Gallegos, J.A. (2013). Phenotyping common beans for adaptation to drought. Front. Physiol. **4**: 35.

Bornowski, N., Song, Q., and Kelly, J.D. (2020). QTL mapping of post-processing color retention in two black bean populations. Züchter Genet. Breed. Res. **133**: 3085–3100.

Broughton, W.J., Hernández, G., Blair, M., Beebe, S., Gepts, P., and Vanderleyden, J. (2003). Beans (*Phaseolus* spp.) – model food legumes. Plant Soil **252**: 55–128.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. **103**: 338–348.

Cichy KA, Porch TG, Beaver JS, Cregan P, Fourie D, Glahn RP, Grusak MA, Kamfwa K, Katuuramu DN, McClean P, et al (2015) A *Phaseolus vulgaris* Diversity Panel for Andean Bean Improvement. Crop Sci **55**: 2149–2160

Costa, J.G.C. da and Rava, C.A. (2003). Linhagens de feijoeiro comum com fenótipos agronômicos favoráveis e resistência ao crestamento bacteriano comum e antracnose. Ciênc. Agrotecnologia **27**: 1176–1182.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al (2011) The variant call format and VCFtools. Bioinformatics **27**: 2156–2158

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience **10**.

Dong L, Fang C, Cheng Q, Su T, Kou K, Kong L, Zhang C, Li H, Hou Z, Zhang Y, et al (2021) Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. Nat Commun **12**: 5445

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6**: e19379.

Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, Hu G, Zhou Z, Yu H, Zhang M, et al (2017) Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. doi: 10.1186/s13059-017-1289-9

Gao, X., Starmer, J., and Martin, E.R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet. Epidemiol. **32**: 361–369.

Gepts, P. and Bliss, F.A. (1985). F1 hybrid weakness in the common bean. J. Hered. **76**: 447–450.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., and Buckler, E.S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One **9**: e90346.

Haghighi, K. and Ascher, P. (1988). Fertile, intermediate hybrids between *Phaseolus vulgaris* and *P. acutifolius* from congruity backcrossing. Sex. Plant Reprod. **1**.

Hannah, M.A., Krämer, K.M., Geffroy, V., Kopka, J., Blair, M.W., Erban, A., Vallejos, C.E., Heyer, A.G., Sanders, F.E.T., Millner, P.A., and Pilbeam, D.J. (2007). Hybrid weakness controlled by the dosage-dependent lethal (DL) gene system in common bean (*Phaseolus vulgaris*) is caused by a shoot-derived inhibitory signal leading to salicylic acid-associated root death. New Phytol. **176**: 537–549.

Harlan, J.R. and Wet, J.M.J. (1971). Toward a rational classification of cultivated plants. Taxon **20**: 509–517.

Hart, J.P. and Griffiths, P.D. (2015). Genotyping-by-sequencing enabled mapping and marker development for the by-2 Potyvirus resistance allele in common bean. Plant Genome **8**: eplantgenome2014.09.0058.

Huang, M., Liu, X., Zhou, Y., Summers, R.M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience **8**.

International Center for Tropical Agriculture (2022). Bean germplasm collection and database. CIAT genebank.

International Commission on Illumination (2008). ISO 11664-4:2008 Colorimetry — Part 4: CIE 1976 L*a*b* Colour space (Geneva, Switzerland).

Jiménez, J.C., de la Fuente, M., Ordás, B., García Domínguez, L.E., and Malvar, R.A. (2017). Resistance categories to *Acanthoscelides obtectus* (Coleoptera: Bruchidae) in tepary bean (*Phaseolus acutifolius*), new sources of resistance for dry bean (*Phaseolus vulgaris*) breeding. Crop Prot. **98**: 255–266.

Kamfwa, K., Cichy, K.A., and Kelly, J.D. (2015). Genome-wide association analysis of symbiotic nitrogen fixation in common bean. Züchter Genet. Breed. Res. **128**: 1999–2017.

Katuuramu, D.N., Hart, J.P., Porch, T.G., Grusak, M.A., Glahn, R.P., and Cichy, K.A. (2018). Genome-wide association analysis of nutritional composition-related traits and iron bioavailability in cooked dry beans (*Phaseolus vulgaris* L.). Mol. Breed. **38**.

Kornegay, J.L. and Cardona, C. (1991). Inheritance of resistance to *Acanthoscelides obtectus* in a wild common bean accession crossed to commercial bean cultivars. Euphytica **52**: 103–111.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Liu, X., Huang, M., Fan, B., Buckler, E.S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient Genome-wide association studies. PLoS Genet. **12**: e1005767.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M, et al (2020) Pan-genome of wild and cultivated soybeans. Cell **182**: 162-176.e13

Mbogo, K.P., Davis, J., and Myers, J.R. (2009). Transfer of the Arcelin-Phytohaemagglutinin-Î± Amylase Inhibitor Seed Protein Locus from Tepary bean (*Phaseolus acutifolius* A. Gray) to Common Bean (*P. vulgaris* L.). Biotechnology (Faisalabad) **8**: 285–295.

McClean, P.E., Bett, K.E., Stonehouse, R., Lee, R., Pflieger, S., Moghaddam, S.M., Geffroy, V., Miklas, P., and Mamidi, S. (2018). White seed color in common bean (*Phaseolus vulgaris*) results from convergent evolution in the P (pigment) gene. New Phytol. **219**: 1112–1123.

McClean, P.E., Burridge, J., Beebe, S., Rao, I.M., and Porch, T.G. (2011). Crop improvement in the era of climate change: an integrated, multi-disciplinary approach for common bean (*Phaseolus vulgaris*). Funct. Plant Biol. **38**: 927–933.

Mejía-Jiménez, A., Muñoz, C., Jacobsen, H.J., Roca, W.M., and Singh, S.P. (1994). Interspecific hybridization between common and tepary beans: increased hybrid embryo growth, fertility, and efficiency of hybridization through recurrent and congruity backcrossing. Züchter Genet. Breed. Res. **88**: 324–331.

Miklas, P.N. and Santiago, J. (1996). Reaction of select tepary bean to bean golden mosaic virus. HortScience **31**: 430–432.

Miklas, P.N., Schwartz, H.F., Salgado, M.O., Nina, R., and Beaver, J. (1998). Reaction of Select Tepary Bean to Ashy Stem Blight and Fusarium Wilt. hortscien **33**.

Moghaddam SM, Mamidi S, Osorno JM, Lee R, Brick M, Kelly J, Miklas P, Urrea C, Song Q, Cregan P, et al (2016) Genome-wide association study identifies candidate loci

underlying agronomic traits in a Middle American Diversity Panel of common bean. Plant Genome. doi: 10.3835/plantgenome2016.02.0012

Moghaddam SM, Oladzad A, Koh C, Ramsay L, Hart JP, Mamidi S, Hoopes G, Sreedasyam A, Wiersma A, Zhao D, et al (2021) The tepary bean genome provides insight into evolution and domestication under heat stress. Nat Commun **12**: 2638

Mwale, S.E., Shimelis, H., Mafongoya, P., and Mashilo, J. (2020). Breeding tepary bean (*Phaseolus acutifolius*) for drought adaptation: A review. Plant Breed. **139**: 821–833.

Myers, J.R. and Kusolwa, P.M. (2011). Seed storage proteins ARL2 and its variants from the APA locus of wild tepary bean G40199 confers resistance to *Acanthoscellides obtectus* when expressed in common beans. African Crop **19**: 255–265.

Pastor-Corrales, M.A., Steadman, J.R., Urrea, C.A., Blair, M.W., and Venegas, J.P. (2011). The domesticated tepary bean accession G40022 has broader resistance to the highly variable bean rust pathogen than the known rust resistance genes in common bean. Annual report of the bean **54**: 124–125.

Porch, T., Barrera, S., Berny Mier y Teran, J.C., Díaz-Ramírez, J., Pastor-Corrales, M., Gepts, P., Urrea, C.A., and Rosas, J.C. (2022). Release of tepary bean TARS□Tep 23 germplasm with broad abiotic stress tolerance and rust and common bacterial blight resistance. J. Plant Regist. **16**: 109–119.

Porch, T., Beaver, J., Debouck, D., Jackson, S., Kelly, J., and Dempewolf, H. (2013a). Use of wild relatives and closely related species to adapt common bean to climate change. Agronomy (Basel) **3**: 433–461.

Porch, T.G., Beaver, J.S., and Brick, M.A. (2013b). Registration of tepary germplasm with multiple□stress tolerance, TARS□tep 22 and TARS□tep 32. J. Plant Regist. **7**: 358–364.

Raj, A., Stephens, M., and Pritchard, J.K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics **197**: 573–589.

Ramirez-Cabral, N.Y.Z., Kumar, L., and Taylor, S. (2016). Crop niche modeling projects major shifts in common bean growing areas. Agric. For. Meteorol. **218–219**: 102–113.

Rendón-Anaya M, Montero-Vargas JM, Saburido-Álvarez S, Vlasova A, Capella-Gutierrez S, Ordaz-Ortiz JJ, Aguilar OM, Vianello-Brondani RP, Santalla M, Delaye L, et al (2017) Genomic history of the origin and domestication of common bean unveils its closest sister species. Genome Biol. doi: 10.1186/s13059-017-1190-6

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet **46**: 707–713

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature **463**: 178–183

Scott, M.E. and Michaels, T.E. (1992). Xanthomonas resistance of *Phaseolus* interspecific cross selections confirmed by field performance. HortScience **27**: 348–350.

Sedivy, E.J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol. **214**: 539–553.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. **44**: 825–830.

Singh, S.P. and Ariel Gutiérrez, J. (1984). Geographical distribution of the DL1 and DL2 genes causing hybrid dwarfism in *Phaseolus vulgaris* L., their association with seed size, and their significance to breeding. Euphytica **33**: 337–345.

Singh, S.P. and Muñoz, C.G. (1999). Resistance to common bacterial blight among *Phaseolus* species and common bean improvement. Crop Sci. **39**: 80–89.

Soltani A, MafiMoghaddam S, Oladzad-Abbasabadi A, Walter K, Kearns PJ, Vasquez-Guzman J, Mamidi S, Lee R, Shade AL, Jacobs JL, et al (2018) Genetic analysis of flooding tolerance in an Andean diversity panel of dry bean (*Phaseolus vulgaris* L.). Front Plant Sci. doi: 10.3389/fpls.2018.00767

Teiwes, H. and Nabhan, G.P. (1983). Tepary Beans, O'odham Farmers, and Desert Fields. Desert Plants **5**.

Thomas, C.V. and Waines, J.G. (1984). Fertile backcross and allotetraploid plants from crosses between tepary beans and common beans. J. Hered. **75**: 93–98.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. **91**: 4414–4423.

Vlasova A, Capella-Gutiérrez S, Rendón-Anaya M, Hernández-Oñate M, Minoche AE, Erb I, Câmara F, Prieto-Barja P, Corvelo A, Sanseverino W, et al (2016) Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. Genome Biol **17**: 32

Wang, J. and Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. Genomics Proteomics Bioinformatics **19**: 629–640.

Zuiderveen, G.H., Padder, B.A., Kamfwa, K., Song, Q., and Kelly, J.D. (2016). Genome-Wide Association Study of Anthracnose Resistance in Andean Beans (*Phaseolus vulgaris*). PLoS One **11**: e0156391.