ACCELERATING PLANT BREEDING THROUGH DISEASE RESISTANCE SCREENING, PHENOLIC COMPOUND PROFILING, AND SPECTRAL MODELING

By

Sidney Connor Sitar

A THESIS

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

Plant Breeding, Genetics, and Biotechnology – Crop and Soil Sciences – Master of Science

2023

ABSTRACT

Advancing crop development is essential to meet the growing demand for agricultural products, particularly for combating stressors and improving nutritional value. Here, we present a quantitative trait loci (QTL) analysis on tar spot of maize, a genome wide association study (GWAS) on phenolic compound accumulation in maize kernels, and Fourier transform-infrared (FT-IR) spectroscopy used to model the phenolic compounds previously found in kernels. This research maps the genetic resistance to maize tar spot disease using a structured QTL analysis on a Stiff-Stalk MAGIC population. The use of the structured population with multi-location field trails resulted in multiple significant QTL that can be used for candidate gene extraction and the future breeding of resistant maize varieties. Additionally, this research focuses on the phenolic compound profiles of maize kernels to detect the natural range of accumulation and to understand the genetic architecture of the compounds and their pathways. GWAS highlighted 170 significant SNPs and 390 potential candidate genes that contribute to phenolic compound accumulation in maize kernel tissues. Lastly, this research uses FT-IR spectroscopy to predict the phenolic compound accumulation in maize kernels without the costly analytical chemistry. Random forest and partial least squares regression modeling types and numerous spectral preprocessing techniques were tested for their accuracy to model these phenotypes. These studies contribute to crop improvement by providing tools that can be used in future plant breeding programs that promote disease resistance, increased nutritional elements, and rapid phenotyping to create more efficient pipelines.

This thesis is dedicated to my family and friends. Thank you for all your continued support throughout my journey. This would not have been possible without you.

ACKNOWLEDGEMENTS

First, I would like to express my deepest appreciation to my advisor Dr. Addie Thompson for her continued support and mentoring throughout my journey at Michigan State University. I appreciate all the meaningful advice and opportunities you have given me as a member of your lab. I am also extremely grateful for the passion for maize and coding you have passed down to me. I would also like to thank Linsey Newton for the assistance at each and every stage throughout my time here at Michigan State University. I am appreciative of the field and laboratory skills you have passed on to me throughout my graduate and undergraduate career. I would also like to thank the members of my committee, Dr. Martin Chilvers and Dr. Erich Grotewold for the guidance in my research. I am appreciative of the interdisciplinary expertise you have brought to my project. Finally, I would also like to acknowledge the members of the Thompson maize genetics lab who were essential to the progression of my research. I appreciate the meaningful contribution to my data collection, the insightful comments, and friendships along the way.

TABLE OF CONTENTS

CHAPTER 1: LITERATURE REVIEW OF TAR SPOT OF MAIZE, PHENOLIC COMPOUND ANALYSIS, AND FT-IR SPECTROSCOPY APPLICATIONS FOR PLANT BREEDING PURPOSES

ABSTRACT

Plant breeding is a multidisciplinary field that draws on various domains such as genetics, phenomics, biochemistry, pathology, and agronomy. Its progress spans centuries, driven by the need to develop resilient crops capable of thriving in diverse stressful and changing environments. Research innovations are essential to further enhance the efficiency of traditional plant breeding. Among the main threats to crop production are diseases and other biotic or environmental stressors. Plant diseases cause devastating effects to yield when not managed properly. One relatively recent concern in the United States is tar spot disease in maize, characterized by the emergence of small black lesions on the foliar tissue. As this fungal disease progresses, it can cause leaf senescence, tissue necrosis, premature stalk lodging, reduced grain fill, and, ultimately, a substantial reduction in yield. Enhancing genetic resistance to diseases and fortifying the plant's defense system is essential to mitigate these yield losses. Phenolic compounds and flavonoids in maize play pivotal roles in the plant's defense response and immune system, in addition to providing numerous health benefits. Phenotyping of phenolic compounds is traditionally a labor-intensive, costly process requiring substantial analytical chemistry. Exploring techniques like Fourier transform – infrared spectroscopy could lead to rapid phenolic compound detection and quantification without the expenses of analytical chemistry. Leveraging insights from genetic disease resistance, phenolic compound accumulation, and FT-IR spectroscopy detection methods will improve the plant breeding process, resulting in the development of more elite crop varieties.

PART 1 – GENETIC RESISTANCE TO TAR SPOT OF MAIZE

Fungal pathogen causing tar spot disease: Phyllachora maydis

Tar spot of maize is caused by the fungal pathogen *Phyllachora maydis. P. maydis* is an obligate ascomycete fungus. Ascomycota is the largest fungal phylum with many of the species used as model organisms for genetic research (Naranjo-Ortiz & Gabaldón, 2019). Ascomycota have a wide range of fungal types, ranging from simple yeasts to fungi with complex macroscopic fruiting bodies (Naranjo-Ortiz & Gabaldón, 2019). Ascomycete fungi are known as sac fungi, named after the specialized asci, or sacs, that hold the sexual spores (Bennett & Turgeon, 2016). Obligate biotrophs require live plant material to survive and complete their life cycle, including the reproduction of additional spores (Cannon, 1991).

P. maydis produces two spore types throughout its lifecycle. The disease cycle is not completely understood, but some parts are known. This disease has a polycyclic cycle, meaning there are multiple rounds of inoculum produced, so infection can occur multiple times throughout the same growing season (Bajet et al., 1994; Hock et al., 1989). The primary inoculum for *P. maydis* infection is thought to be infected residue from the previous field season (Groves et al., 2020), or spores blown in from other nearby fields (Hock et al., 1992). According to research by Hock et al. (1992), ascospores, the sexual spores, are released from the stromata and travel through wind and rain splash to nearby maize plants. *P. maydis* thrives in environmental conditions with high leaf wetness, high humidity, and moderate temperature (Hock et al., 1995; Valle-Torres et al., 2020). At the end of the growing season, the ascospores and conidia of *P. maydis* can overwinter in stromata on decaying plant residue (Groves et al., 2020). This makes controlling the disease and infection especially difficult with current management practices.

History of tar spot of maize

P. maydis was first identified in Mexico in 1904 and was endemic to Latin America for many years (Bajet et al., 1994; Liu, 1973). Tar spot of maize has been identified in most Central American regions, the Caribbean, and multiple South American countries (Hock et al., 1989). In Mexico, yield losses of up to 58% have been recorded in susceptible maize hybrids due to tar spot (Loladze et al., 2019). It was once thought that tar spot of maize was confined to tropical regions, but in recent years the disease has made appearances in new locations. Although tar spot is not a new disease in itself, it is new to the United States. *P. maydis* was first discovered in the United States in 2015 in fields in Illinois and northern Indiana (Ruhl et al., 2016 Since then, it has quickly spread throughout the majority of the Midwest and scattered throughout other areas of the United States (Corn ipmPIPE, 2023; Ruhl et al., 2016). In 2018, the United States experienced especially severe cases of tar spot, with the devastating effects impacting the crop yields. Tar spot led to an almost 5 million metric tons reduction in yield, resulting in economic losses exceeding 680 million USD (Mueller et al., 2020). Similarly, the United States experienced another year of severe tar spot disease in 2021. Estimates suggest that this disease caused a grain yield loss of 5.88 million metric tons, translating to an economic loss of 1.25 billion USD (Crop Protection Network, 2023).

Effects of tar spot of maize

Tar spot of maize is identified by the formation of dense black stromata on the leaves and husks of maize. The black lesions resemble small bits of tar and the density of the lesions can vary from light to heavy, depending on the severity. One of the primary ways to distinguish tar spot lesions from other similar looking occurrences is that the stromata does not rub off with

applied pressure, and the lesions go all the way through the leaf (Telenko & Creswell, 2019). If the primary inoculum for *P. maydis* infection is from the overwintering spores on plant residue, symptoms normally start with the lower leaves, then spread to the rest of the foliage (Cline, 2019). If the spores travel by wind or water splash, the symptoms can be found on upper leaves first (Cline, 2019). In some cases, the black stromata are encased by brown necrotic halo type lesions (fish-eye lesions) that add to the rapid necrosis of the tissue, especially in Mexico (Cline, 2005). Over time, the halos, created by the fish-eye lesions surrounding the black stromata, coalesce and the entirety of the leaf becomes chlorotic and necrotic (Yan et al., 2022). This can lead to rapid canopy senescence in severe infections (Telenko et al., 2019).

Yield loss is a significant concern associated with tar spot infection. The extent of loss can vary depending on factors such as the timing of infection, environmental conditions, and the susceptibility of the hybrid (Telenko & Creswell, 2019). Yield loss occurs due to reduced ear weight, poor kernel fill, and vivipary (Telenko et al., 2020). There is also an increase in stalk rot and lodging when the disease severity is high (Telenko et al., 2020). In extreme cases, tar spot can also reduce silage corn feed quality by reducing the moisture, digestible components, and energy (Telenko et al., 2020). Although tar spot has only been in the United States for a few years, it has led to devastating economic losses due to its significant impact on agriculture.

Current management practices of tar spot

Current management practices of tar spot can help limit the effects of disease, but more work is needed to find a cost effective, environmentally friendly method that eliminates the impact of disease. Farmers have limited options for successful management of tar spot of maize. One of the first control points is with the hybrid seed they choose to plant, as it is best for the

farmer to avoid planting hybrids with high susceptibility to disease (Telenko et al., 2020). Avoiding the highly susceptible hybrids is crucial for reducing disease incidence and lowering overall disease pressure (Telenko et al., 2020). Another option for control is fungicide application. Fungicides are popular for lowering the level of disease within a field, but application strategies are sensitive to timing for optimal success. Application strategies are dependent on disease variability and environmental conditions, which change every growing season (Telenko et al., 2020). Although fungicides are deemed very useful, there are still downsides to dependence on them. Fungicides can be harmful to the environment, costly, require specialized equipment, and require special certifications for application.

Understanding the conducive conditions of tar spot infection and spread are also important for disease management. Tar spot is known to thrive in humid and wet conditions, where leaf wetness is extended (Hock et al., 1995). The overuse of irrigation can create this optimal environment for *P. maydis* due to the increase in duration of leaf wetness and humidity. Reducing the frequency and duration of leaf wetness through limiting or strategically timing irrigation can be a useful strategy in reducing the optimal conditions for infection and spread (Telenko et al., 2020).

There are also other cultural practices that may limit disease by reducing inoculum. *P. maydis* produces overwintering fungal bodies that survive on decaying plant residue in fields; proper residual management through tilling after harvest can promote decomposition of plant tissue that the spores need to survive on. This practice is known for reducing the primary inoculum for future growing seasons (Groves et al., 2020; Telenko et al., 2020). Crop rotation to non-host plants (crops other than maize) also reduces the primary inoculum for future growing seasons, by allowing the plant residue to decompose before maize can be planted in the field

again (Telenko et al., 2020). Lastly, knowledge of the history of the field can be helpful for estimating the diseases that will be present in current growing seasons. Fields with known presence of *P. maydis* will allow the farmer insight on which crop to plant or disease management program to employ specific to the field; as of now, prevention and avoidance are currently the best way to manage tar spot of maize (Telenko et al., 2020).

The current disease management practices have a lot of room for improvement. Even with the above listed techniques, the presence of tar spot can devastate fields for years and pose risk to surrounding fields. Finding genetic resistance in maize will prevent the disease from finding susceptible tissue to establish disease on. It will also reduce the need for costly and potentially harmful fungicide applications.

QTL Analysis

Quantitative trait loci analysis is a statistical approach that links phenotypic trait data measurements with genotypic data (or molecular markers) with the goal to explain the variation in the complex phenotypic traits with the genotypic information (Falconer & Mackay, 1996). One of the main questions QTL analyses tries to answer is whether the observed phenotypic variation in the population of interest is due to many loci with small effects, or few loci with large effects (Roff, 2007). The data needed for completing a QTL analysis are genetic markers and quantitative trait values. The marker data is considered categorical, and it can come in many forms, depending on the population, type of analysis, and method (Zeng, 2001). The primary objective of QTL analysis is to establish a connection between quantitative trait variation and genetic marker variation, utilizing a genetic model that incorporates various genetic architecture features (Zeng, 2001).

Although QTL analysis is powerful, there are still limitations to the method. Successful QTL analysis needs large sample sizes to have enough power to detect small differences in phenotypes (Miles & Wayne, 2008). Another caveat to QTL analysis is some loci of significance will not be uncovered (Miles & Wayne, 2008). Many complex quantitative traits are controlled by numerous loci with very minute effects, which can be difficult to detect. Also, QTL analysis only provides a statistical method to locate the loci of significance; other techniques are still needed to identify the causative gene. This can be especially challenging in populations with a low number of recombination events, as mapping resolution will be low. Additionally, there is more research needed to connect the function of the candidate gene with the QTL discovered through the analysis. This can be time consuming and requires many different downstream techniques to validate (Miles & Wayne, 2008). Lastly, the resolution of the genetic map used for QTL analysis will also contribute to the power of the analysis, and the ability to find peaks of interest within a small area of the genome (Miles & Wayne, 2008). Inadequate marker density leads to lowered power for the statistical method.

There are multiple different types of QTL analysis that are suited to different statistical power levels, data types, and goals.

One marker analysis:

One marker analysis is the simplest of the methods for associating the genetic marker with phenotypic trait variation (Zeng, 2001). One marker analysis uses a *t* test to test for significant difference between the trait means for groups of individuals with the marker genotypes. If the means are found to be significantly different, the marker is determined to be linked to one or more QTL (Zeng, 2001). The downfall of this method is that it cannot be determined if it is one or more QTL the marker is linked to and

depending on map density, it can be difficult to determine how closely linked the marker is with the QTL (Zeng, 2001).

Interval mapping:

Another analysis method is interval mapping. Lander and Botstein (1989) developed a maximum likelihood method that tests an interval of a chromosome relative to its flanking markers (Zeng, 2001). This method is helpful for evaluation at marker locations and between them for a better estimation of QTL position. A downfall of interval mapping is results can be biased if there are multiple QTL on a single chromosome (Zeng, 2001).

Composite interval mapping:

An alternative QTL method is composite interval mapping. Composite interval mapping uses other markers as cofactors to increase the success in estimation of genetic background interactions and to reduce the bias caused by multiple QTL linked to the interval of interest (Bernardo, 2020). This method combines interval mapping and multiple regression techniques to individually test each interval for a QTL (Zeng, 2001). However, there are caveats to composite interval mapping. The analysis can be impacted by an uneven distribution of markers in a genome, which may cause the test statistics for different regions to be incomparable. In addition, it is difficult to measure epistasis of multiple QTL, and to estimate the contribution of multiple QTL to the phenotypic variance (Zeng, 2001).

Multiple interval mapping:

Another technique, multiple interval mapping, looks to address the limitations of composite interval mapping. Multiple interval mapping accomplishes this with fitting

multiple QTL, including epistasis effects, in a model to search, test, and estimate the positions, effects, and interactions of multiple QTL simultaneously (Zeng, 2001). Multiple interval mapping has four main components: first, an evaluation protocol to analyze the likelihood of the data with a genetic model; then, a search strategy to select the best genetic model; next, an estimation procedure for all genetic parameters of the quantitative traits (number, location, effects, epistasis, genetic variance, and covariance) explained by QTL from the selected genetic model. Lastly, a prediction method is employed to estimate the genotypic values of individuals and the offspring based on the genetic model and the estimated genetic parameter values for marker-assisted selection (Kao et al., 1999; Zeng et al., 1999; Zeng, 2001). A downfall of multiple interval mapping is the requirement to identify a subset of markers that account for the QTL in other locations of the genome (Bernardo, 2020).

Populations for QTL analysis

F2 Mapping populations

F2 mapping populations are created by crossing two distinct lines, followed by self-pollination of the resulting F1 progeny. These populations offer the advantage of being easy and rapid to establish. However, they are constrained by the relatively low number of recombinations that can occur in the limited meiotic events. Each F2 population can be cultivated for only one season in the case of annual species, requiring the recreation of the population for future studies involving those parents, including the phenotyping and genotyping of each individual once again. Additionally, the success of mapping is confined to the genetic variation present in the two parents.

Recombinant Inbred Lines

Recombinant inbred line populations overcome some of the shortcomings of F2 populations in that they are "eternal" since they are fully inbred and can be regenerated for repeated measurements. To develop these populations, two contrasting parents are crossed, and the resulting segregating population is self-pollinated several generations until reaching near total homozygosity. They are still limited by the number of recombinations, as well as only containing the alleles present in the two founding parental lines. However, they are very powerful for genetic mapping

MAGIC populations

Multi-parent Advanced Generation InterCrosses (MAGIC) breeding designs are deemed useful in modern plant breeding and research practices. MAGIC populations are considered a multi-parent cross design with the function of creating panels for recombinant inbred lines (RIL) that are mosaics of the founding parents' genomes (Dell'Acqua et al., 2015). Collecting data with a population with multiple founders creates genetic tools that are beneficial for mapping, have high power and resolution for detecting quantitative trait loci, and have a high genetic diversity (Scott et al., 2020). Another benefit of using MAGIC populations is the production of a reusable reference population that allows phenotypic data to be collected over multiple growing seasons, with reduced genetic mapping costs (Dell'Acqua et al., 2015). Studying MAGIC populations for phenotypic traits, such as disease resistance traits, can uncover new genes of interest and validate previously uncovered genes.

The MAGIC population used for this study is made up of 6 inbred lines thought to represent the range of diversity in the stiff stalk heterotic pool (Michel et al., 2022). The 6

stiff stalk founder inbreds are B73, B84, NKH8431, LH145, PHB47, and PHJ40 (Michel et al., 2022). The founders B73 and B84 originated at Iowa State University, and both are a part of the B73 sub-heterotic group (White et al., 2020). The founding lines LH145 and NKH8431 originated from Holden's Foundation Seed, Inc. and Northrup, King & Company respectively, and are a part of the B14 sub-heterotic group (White et al., 2020). PHB47 and PHJ40 originated from Pioneer Hi-Bred International, Inc. and are a part of the B37 and Flint sub-heterotic group respectively (White et al., 2020). 500 lines were derived from the crossing scheme with the 6 founders as parents, given the population the name of WI-SS-MAGIC (Michel et al., 2022). This structured MAGIC population allowed for a powerful QTL mapping structure due to the increased genetic diversity that the additional founders provide (Scott et al., 2020).

Previous research on genetic resistance to tar spot

With the current disease management practices for tar spot of maize lacking, especially with the increased spread and devastating yield effects, there is need for additional methods of resistance. Genetic resistance to disease is a powerful tool for breeding non susceptible hybrids. It is a cost-effective method for growers that reduces the need for costly fungicides and equipment. Genetic resistance to tar spot of maize has started to be studied by a few research groups.

One research study looked at a total of 890 maize inbred lines, chosen to broadly represent tropical and subtropical maize genetic diversity, from CIMMYT maize lines and varieties from breeding programs researching stressor resistance, grown over multiple years and environments (Mahuku et al., 2016). Mahuku et al. (2016) used low density markers for QTL

mapping in the bi-parental populations with an inclusive composite interval mapping technique. These methods uncovered a major QTL (*qRtsc8-1*) connected with resistance to tar spot of maize and occurring at a frequency of 3.5% the total maize varieties they studied (Mahuku et al., 2016).

Another research study aimed to dissect the genetic architecture of tar spot resistance in maize through association mapping with linkage mapping, using an association mapping panel and three biparental doubled-haploid (DH) populations (Cao et al., 2017). The association-mapping panel Cao et al. (2017) used was originally designated as the Drought Tolerant Maize for Africa (DTMA). The association mapping in this study uncovered 4 QTL on chromosome 2, 3, 7, and 8, and the linkage mapping validated all of the QTL except the one on chromosome 3 (Cao et al., 2017). The QTL located on chromosome 8 (bin 8.03), was consistently detected and explained the largest phenotypic variation and concluded that this major QTL along with several minor small effect QTL controlled the disease resistance in this study (Cao et al., 2017).

Another study, performed by Yan et al. (2022), looked at an association mapping panel of 228 CIMMYT maize lines developed from the different breeding programs for superior yield, quality, and stressor resistance performance. A genome wide association study (GWAS) was completed using a mixed linear modeling (MLM) approach with 5 principal components to detect potential SNPs of significance (Yan et al., 2022). The GWAS done by Yan et al. (2022) discovered 178 significantly associated SNPs that were distributed in five QTL regions. Six of the SNPs located in bins 2.02, 3.03, 3.06, and 10.05 (Yan et al., 2022). The remaining 172 significant SNPs located in bin 8.03 (Yan et al., 2022). The research done by Makuku et al. (2016), Cao et al., (2017), and Yan et al. (2022) found a significant QTL located in bin 8.03.

With the multiple studies finding significant genetic resistance located on chromosome 8 at bin 8.03, another research study, Ren et al. (2022), decided to use fine mapping for the major

QTL for tar spot resistance to verify the effects of the markers and to speed up the development of breeding lines with tar spot resistance (Makuku et al., 2016; Cao et al., 2017; Yan et al., 2022). Fine mapping is a technique that will allow the verification of the maker significance for future creation of elite breeding material through marker assisted selection (Badu-Apraku & Fakorede, 2017). Ren et al. (2022) fine mapped *qRtsc8-1* with an interval of 721 kb flanked by the markers KASP81160138 and KASP81881276 in the BC_5 generation. At this interval, the two candidate genes *GRMZM2G063511* and *GRMZM2G073884* were identified (Ren et al., 2022). The gene *GRMZM2G063511* encodes an integral membrane protein-like, and the gene *GRMZM2G073884* encodes a leucine-rich repeat receptor-like protein kinase (Ren et al., 2022). There has been other research done searching gray leaf spot and common rust resistance in maize that have found significant QTLs or SNPs in bin 8.03 (Benson et al., 2015; Mammadov et al., 2015; Shi et al., 2014; Olukolu et al., 2016; Zheng et al., 2018). It is likely that both genes are involved in maize disease resistance response and can be used in the future for breeding elite varieties with strong resistance to tar spot.

Another research study looked at over 600 varieties from the Wisconsin Diversity panel and 200 varieties from Iowa State's Germplasm Enhancement of Maize program to screen for tar spot resistance (Trygestad, 2021). This study performed a GWAS using the Genome Association and Prediction Integrated Tool (GAPIT) package in R (Lipka et al., 2012) with the fixed and random model Circulating Probability Unification (FarmCPU) method (Liu et al., 2016) to uncover significant SNPs (Trygestad et al., unpublished). The research uncovered over 100 significant SNPs connected with tar spot resistance and linked with candidate genes. None of the significant SNPs found by Trygestad et al. (unpublished) were previously identified in tropical maize germplasm by Cao et al. (2017).

Validation of the results found by previous researchers (Makuha et al., 2016; Cao et al., 2017; Yan et al., 2022; Ren et al., 2022, Trygestad et al., unpublished) will provide additional confidence for incorporating candidate genes into new elite maize varieties. Additional research could increase the number of minor and major QTLs found all leading to tar spot resistance. An experiment mapping tar spot resistance in a MAGIC population could provide an untested technique for uncovering genetic resistance to tar spot. This could lead to a higher power study that is able to detect more QTL, validate previous results, and evaluate the founders of the MAGIC population to see which parent contributed the most resistance alleles in which part of the genome.

PART 2 - PHENOLIC COMPOUND DETECTION AND ACCUMULATION IN PLANTS

Introduction on phenolic compounds

Phenolics are chemically identified as compounds that contain a hydroxylated aromatic ring, with the hydroxy group attached directly to the phenyl, substituted phenyl, or aryl group (Swanson, 2003). Phenolic compounds are considered specialized metabolites that are derived from either phenylalanine or tyrosine (Shahidi & Naczk, 2003). There are thousands of known compounds and these compounds are distributed generously through plant tissues and are known to greatly contribute to color, flavor, and astringency of plants (Swanson, 2003). Phenolic compounds are classified into groups including phenols, coumarins, lignins, lignans, condensed and hydrolysable tannins, phenolic acids, and flavonoids (Soto-Vaca et al., 2021). Flavonoids can then be separated into smaller sub-groups based on structural differences, these groups include anthocyanins, flavonols, flavones, flavan-3-ols, isoflavones, and flavanones (Šamec, et al., 2021). Together, flavonoids and phenolic acids make up the majority of dietary phenolic

compounds. Flavonoids are highly abundant in the majority of fruits and vegetables, although, the type and concentration of flavonoids vary based on the plant and the tissue (Xu et al., 2017; Erlund, 2004).

Biosynthesis and metabolism of phenolic compounds in plants

Phenolic compounds, and specifically flavonoids, have important roles in plant metabolism and biology, and the biosynthesis pathways are greatly studied. Patterns of secondary metabolites in plants are complex because they change between tissue types and they evolve as the plant goes through the different developmental stages (Lattanzio et al., 2012). The metabolism of phenolics in plants includes numerous biosynthesis pathways and processes (Lattanzio et al., 2012). The products from the shikimate pathway, either phenylalanine or tyrosine, are used in the first steps of the phenylpropanoid pathway. Phenylalanine is converted to cinnamic acid, which is the precursor to the conversion to creation of the other phenolic acids (Cocuran et al., 2019). Then, the phenylpropanoid pathway produces *p*-coumaryl-CoA, which is used as the precursor for either flavonoid biosynthesis or additional phenolics and lignin precursor formation (Falcone Ferreyra et al., 2012; Cocuran et al., 2019). The flavonoid pathway uses the specific enzyme chalcone synthase to produce chalcone scaffolds; the chalcone scaffolds are a molecule that all flavonoids are derived from (Falcone Ferreyra et al., 2012). Next, a group of enzymes, the type is dependent to the plant species, create the central pathway backbone for flavonoid biosynthesis, leading to the different subgroups (Martens et al., 2010). Finally, transferase enzymes modify the flavonoid backbone with different molecules including sugars, methyl groups, and acyl moieties (Bowles et al., 2005). These modifications to the flavonoid backbones determine the physiological activity and uses for the resulting flavonoid created

(Ferrer et al., 2008).

Historically, there has been many research inquiries and immense interest in deciphering these biosynthesis pathways through a genetic perspective. Maize (*Zea mays*) was one of the first model species that was used experimentally to isolate many structural and regulatory flavonoid genes (reviewed in Mol et al., 1998). Determining the genetic control behind the structural and regulatory genes has allowed the understanding of the numerous roles phenolic compounds play within plants, and the diversity of the roles between plants.

Role of phenolic compounds in the plant

Flavonoids are important components of plant metabolism, defense and immune response (Liu et al., 2013). Numerous studies have demonstrated that the intermediate metabolites formed during the biosynthesis of flavonoids are essential to plant physiological metabolic processes (Liu et al., 2013). Additionally, flavonoids are responsible for protecting cereal crops against numerous biotic and abiotic stressors including UV protection, insect resistance, disease resistance, developmental functions, and auxin regulation (summarized in Liu et al., 2013). These important compounds are also essential for plant adaptation in new habitats (Bais et al., 2003) and providing the resources for many of the steps in highly successful reproduction (Dudareva et al., 2004). Flavonoid biosynthesis and the intermediate metabolites created are essential to different processes that occur throughout the entire lifecycle of the plant. Phenolic compounds, mainly flavonoids, are attributed to seed dormancy, hormone regulation, root nodule development, and bacterial signaling (summarized in Liu et al., 2013). Flavonoids are also known to absorb UV rays which provides the plant with UV radiation protection needed for response to stress caused by an overabundance of light (Bashandy et al. 2009). Flavonoids and

other key phenolic compounds provide essential roles within the plant that are key for survival and reproduction. There is a lot of research interest on the roles of flavonoids in the plant, but there is still more to be discovered. Additional research studies on how individual flavonoids and phenolic compounds explain phenotypic variation in plants could provide insight on other important roles of these compounds in plants.

Health benefits of consuming phenolic compounds

In addition to the numerous functions and important roles phenolic compounds have within plants, there are also astounding health benefits phenolic compounds have when consumed. Phenolic compounds, including flavonoids, are readily found in fruits and vegetables, and are thought to contribute to the many health benefits from a diet rich in fruit and vegetables (Ballard & Junior, 2019). Although the mechanisms of action of flavonoids are of interest to many researchers, some of them are still not fully understood. Flavonoids, either individually or in combination, have showed antioxidant, anti-inflammatory, anti-diabetic, anti-cancer, anti-obesity, and cardioprotective effects both in invitro and in vivo models (Xiao et al., 2011; reviewed in Ballard & Junior, 2019). It is thought that under oxidative stress, polyphenols participate in modulation of redox status and with intracellular signal transduction pathways related to cell proliferation, apoptosis, inflammation, angiogenesis, and metastasis (Basli et al., 2017).

Multiple research studies, reviewed in Basli et al. (2017) showed evidence of the correlation of phenolic compounds slowing down, or in some cases preventing, the initiation/progression of multiple types of cancers. The antioxidant properties of flavonoids piqued interest in their role with cardiovascular health (Kozłowska & Szostak-Węgierek, 2014).

Flavonoid ingestion is also known to help with insulin resistance and type II diabetes mellitus (reviewed in Ballard & Junior, 2019). The benefits of flavonoid ingestion correlate with the reduction of the amount of reactive oxygen species (ROS) while having a direct effect on pro-oxidant enzymes with antidiabetic function (Habtemariam & Varghese, 2014). These effects increase insulin secretion which promotes the proliferation of pancreatic β cells which results in regulation of glucose metabolism (Babu et al., 2013, Ballard & Junior, 2019). The anti-obesity properties of flavonoids are also widely studied. In vitro research studies suggest that flavonoids reduce the viability and proliferation of adipocytes which results in suppressing triglyceride accumulation, reducing inflammation, and stimulating lipolysis (reviewed in Ballard & Junior 2019).

Flavanones, a subclass of the flavonoid group of phenolic compounds, act as antioxidants and play a significant role in anti-inflammatory response (Bredsdorff et al., 2010; Soto-Vaca et al., 2012). Flavonols are one of the most widely distributed plant flavonoids found in human diets and have demonstrated benefits with cardiovascular disease when consumed as fruits or vegetables (Wang et al., 2009). Flavones are normally found in small amounts in herbs, grains, and leafy vegetables and like flavonols, have also been linked to a lower risk of cardiovascular disease (Perez-Vizcaino & Duarte, 2010; Lin et al., 2007). Isoflavones, which are found mostly in legumes like soy, are found to be beneficial for menopausal symptoms in women and the reduction of low-density lipo-protein cholesterol (reviewed in Soto-Vaca et al., 2012). There are many beneficial properties to consuming flavonoid rich foods as a part of the human diet, and increasing intake may cause prevention or slowing of the progression of some life-threatening diseases. More research is always needed to understand more of the mechanisms of phenolic compounds in our bodies, the health benefits they provide, and distinguishing the difference

between correlation and causation of these benefits.

Methods to extract and quantify the accumulation of phenolic compounds in plants

Extracting phenolic compounds, including diverse flavonoids, from plant tissues has always been a costly and time-consuming process. Part of the difficulty detecting and quantifying phenolics stems from the chemical diversity, acylation or glycosylation, and the complex biological pathways (Cocuran et al., 2019). Although, modern techniques and technology advancements have reduced energy and solvent use, increased efficiency, and aligned all processes with environmental regulations (Chaves et al., 2020). There are various techniques and protocols for the extraction of phenolics including pressurized liquid extraction, accelerated solvent extraction, extraction assisted by pulsed electric field, enzyme-assisted extraction, solid-phase extraction, microwave-assisted extraction, supercritical fluid extraction, ultrasound-assisted extraction, or a combination of multiple of the above techniques (Chaves et al., 2020). One of the more popular methods includes using organic solvents (methanol, ethanol, acetonitrile, petroleum ether, or acetone) and water to create a solution that extracts flavonoids and other phenolic compounds from diverse plant tissues (Chaves et al., 2020). This is successful due to the polarity of the molecules. Altering the pH with an acid (ie. formic acid) causes the production of protons that are thought to stabilize any potential free radicals extracted from the plant tissue (Chaves et al., 2020).

Liquid chromatography – mass spectrometry (LC-MS) is an analytical chemistry technique that couples the physical separation achieved through liquid chromatography with the mass analysis and quantification properties of mass spectrometry (Cocuron et al., 2019). Success has been found coupling HPLC (high pressure liquid chromatography) with a triple quadrupole

in order to quantify specified targeted compounds (Bataglion et al., 2015; Lin et al., 2015). Additional advancements have been made using ultra-high performance liquid chromatography tandem mass spectrometry (UHPLC-MS/MS) to identify and quantify plant flavonoids and other phenolic compounds including phenolic acids, aldehydes, and alcohols, which was previously lacking (Cocuran et al., 2019). Through the research that Cocuron et al. (2019) completed, accurate and rapid testing for a range of phenolic compounds in plant tissues is more attainable.

The use of Waters ACQUITY TQD Tandem Quadrupole Ultra Performance Liquid Chromatography/Tandem Mass Spectrometry (TQ-D UPLC/MS/MS) allows for high speed and sensitivity, coupled with the high selectivity and power (Waters Corporation, Milford, MA, USA). Ultra-performance liquid chromatography (UPLC) is a form of liquid chromatography where narrow-bore columns packed with small particles and high back-pressure mobile phase delivery systems are used (Gruz et al., 2008). This is an improved method when compared to conventional high-performance liquid chromatography (HPLC) due to the improved resolution, shorter retention times, and increased sensitivity (Yu et al., 2006). The UPLC system paired with electrospray ionisation (ESI) tandem mass spectrometry (MS/MS) has been successfully used to analyze plant and food products (Gruz et al., 2008). The use of the TQ-D UPLC/MS/MS system for phenolic compound detection and quantification in maize kernel samples was selected for the increased sensitivity, high resolution, and shorter retention times.

Analysis on flavonoid accumulation in plants

With all of the health benefits and plant responses that flavonoids are a part of, there have been many research studies with the aim of finding which plants have high accumulations of specific flavonoids. Flavonols are a subclass of flavonoids that have a ketone group on position 4

of the C ring and a hydroxy group on position 3 of the C ring (Panche et al., 2016). Fruits and vegetables like onions, kale, lettuce, tomatoes, apples, grapes, and berries have high flavonol content along with tea and red wine (Panche et al., 2016). Quercetin is commonly the most abundantly accumulated flavonol in edible plants, and quercetin is found in the highest concentration in onion (Xu et al., 2017). Flavones are easily found in leaves, flowers, and fruits as glucosides (Panche et al., 2016). The foods that are major sources of flavones are celery, parsley, red peppers, chamomile, and mint (Panche et al., 2016). The major flavones, apigenin and luteolin, are commonly found in red pepper and celery (Xu et al., 2017). Flavanones, also called dihydroflavones, have a saturated C ring and an additional double bond that distinguishes them from flavones (Panche et al., 2016). Flavanones are generally in all citrus fruits, like oranges, lemons, and grapes, and are known for their free radical-scavenging properties (Panche et al., 2016). This class of flavonoids is responsible for the bitter taste of the juices and peels of citrus fruits (Panche et al., 2016). Isoflavonoids are a very large and distinct group that are only found very limitedly in plants and some have even been reported in microbes (Panche et al., 2016; Matthies et al., 2008). Of the limited plants, isoflavonoids are mainly found in soybeans and other various legumes (Xu et al., 2017; Panche et al., 2016). Anthocyanins are found in high concentrations in most edible plants that have a red, purple, or blue coloring. Anthocyanins are known to contribute to the pigments in plants (Xu et al., 2017). Anthocyanins are pigments responsible for colors of plants, flowers, and fruits; they are found mostly in the outer cell layers of highly colored fruits (Panche et al., 2016). The anthocyanin subgroup are the only flavonoids that are ionic, the group of water-soluble pigments are able to serve as a pH indicator of the vacuole, and are in high abundance in nectarine, black beans, and berries (Rosa et al., 2019). Flavon-3-ols, also called dihydroflavonols, flavavonols, or catechins, are the 3-hydroxy

derivatives of the flavanone group (Panche et al., 2016). This is a highly diverse and multi-substituted subgroup of flavonoids that are found generously in bananas, apples, blueberries, peaches, and pears (Panche et al., 2016).

In maize, phenolic compounds have many important roles that have been of interest to researchers. Phenolics are necessary for conditional male fertility, seed coat development, signaling, seed dormancy, and regulation of the transport of phytohormones (Jin et al., 2017). Through the research uncovered by Zhang et al. (2018), flavonoid concentration was confirmed to vary greatly between different tissue types. When looking at pollen, silks, tassel, and kernel tissue, for eriodictyol, luteolin, isoorientin, and maysin, Zhang et al. (2018) found that eriodictyol was highest in pollen, and luteolin is low in all 4 tissues. Zhang et al. (2018) also found that isoorientin was highest in pollen and tassels, lower in silks, and insignificant in kernels, Lastly, Zhang et al. (2018) found maysin was high in silks and tassels, but not kernel tissue. Although some research has been done on the accumulation of flavonoids in maize kernels, there has not been an in-depth study of numerous phenolic compound accumulation in different varieties. For future breeding efforts to manipulate phenolic compound accumulation, more extensive research on more compounds on diverse maize varieties is needed.

Genome Wide Association Study (GWAS)

One of the main ways to differentiate association mapping from linkage mapping is the use of a general population instead of a specifically designed population. Association mapping allows for QTL detection in a wide variety of mapping panels that have the power to exploit a wide range of genetic diversity. Genome wide association studies (GWAS) are used to associate genotypes with phenotypes by testing for differences in allele frequencies of genetic variants

(Uffelmann et al., 2021). This process involves rapidly scanning markers across genomes to find the genetic variations that are connected to the phenotype of interest (National Institute of Health, 2020). For a successful GWAS, high density markers are needed. This method uses linkage disequilibrium (LD) over small regions to calculate marker-trait associations (). Linkage disequilibrium is defined as the nonrandom distribution of linkage phases due to neighboring genetic variants being inherited together or being highly correlated with one another (Cano-Gamez & Trynka, 2020). When compared to linkage mapping, association mapping offers nultiple advantages, including increased mapping resolution, reduced research time, and greater allele number (Yu & Buckler, 2006; Tibbs Cortes et al., 2021).

There are numerous GWAS models available for researchers to determine what matches their research type, data, and computing power best. Over time, the evolution of these GWAS methods has led to higher efficiency and statistical power for greater research results.

General Linear Model:

The general linear model (GLM) method is the simplest of the different GWAS methods. It was initially developed and used to address population structure (Zhang et al., 2010). The GLM method uses the cofactors from the subpopulation assignment groups to correct for population structure (Pritchard et al., 2000). This model uses only the fixed-effect model which is the most computing efficient but does not have a high statistical power (reviewed in Tibbs Cortes et al., 2021).

Genomic Control:

The genomic control method was developed by Devlin & Roeder (1999), and it is the first method created to address population structure. This method uses markers that are unlikely to affect the trait of interest, or null markers, to estimate the effect of

population structure on the test statistic (Devlin & Roeder, 1999; Tibbs Cortes et al., 2021). The information gathered about the population structure is then used to adjust the final *p* value which results in the reduction of false positives (reviewed in Tibbs Cortes et al., 2021).

Structured Association:

The Structured Association (STRUCTURE) method was developed by Pritchard et al. (2000) was developed after the genomic control method. This method uses null markers to define a set of subpopulations within the dataset (Pritchard et al., 2000). Then, each individual in the study is assigned to one or more of the subpopulations and the assigned subpopulation(s) is then used as a cofactor for the association model (reviewed in Tibbs Cortes et al., 2021).

Mixed Linear Model:

The mixed linear model (MLM) method was created to replace the older methods (Yu et al., 2005). The MLM uses population structure and kinship to account for relatedness between individuals (Yu et al., 2005). Population structure is determined by either using a principal component analysis or the STRUCTURE method (Price et al., 2006; Pritchard et al., 2000). In addition to population structure, the kinship matrix is used to estimate the relatedness among individuals using the supplied genotype data (reviewed in Tibbs Cortes et al., 2021). This method allows for a greater control of false positives than the previous methods (Yu et al., 2005).

Efficient Mixed-Model Association:

Previous methods were created to address population structure and relatedness, leading to highly taxing and time-consuming calculations. The efficient mixed-model

association (EMMA) technique was created to increase the computational efficiency of solving the MLM equations (Tibbs Cortes et al., 2021). The EMMA method improved computational speed by eliminating any redundant matrix operations within the likelihood function (Kang et al., 2008; Tibbs Cortes et al., 2021). This method can also be used to calculate the kinship matrix through using identity state to produce a matrix of pairwise genetic similarity among the individuals in the population (Tibbs Cortes et al., 2021).

An alteration of the EMMA method was also produced by Kang et al. (2010), called EMMA expedited, to continue to improve computational efficiency through approximation. This method applies a computational shortcut into the normally time-consuming mixed model calculations (Kang et al., 2010).

Population Parameters Previously Determined:

Similarly, to the EMMA expedited method, population parameters previously determined (P3D) also uses approximation to improve computational speed and efficiency (Zhang et al., 2010). This model estimates the genetic residual components once through the base model before any SNPs are tested (reviewed in Tibbs Cortes et al., 2021). These variance components are then used when calculating all SNP effects (Zhang et al., 2010).

Factored Spectrally Transformed Linear Mixed Models:

The factored spectrally transformed linear mixed models (FaST-LMM) method was specifically developed to improve speed related to solving MLM equations (Lippert et al., 2011). This method rewrites the likelihood function of the MLM in a form that is easier to evaluate, which results in improved efficiency (Lippert et al., 2011; Tibbs Cortes

et al., 2021). This method differs from the approximation methods by not requiring the assumption that the variance parameters are the same across all SNPS, which increases power (Tibbs Cortes et al., 2021). Also, FaST-LMM only uses a small subset of SNPs to estimate kinship between individuals (Lippert et al., 2011).

The factored spectrally transformed linear mixed model select (FaST-LMM-Select) was created with the goal to increase the calculation of the kinship matrix (Listgarten et al., 2012). FaST-LMM-Select differs from FaST-LMM because of the careful selection of SNPs used for the kinship matrix calculation (Listgarten et al., 2012). This method uses creation of genetic similarity matrices with increasing number of SNPs with a goal to minimize the genomic control factor (Lisgarten et al., 2012). This method has a high computational efficiency when compared to similar methods (Tibbs Cortes et al., 2021).

Genome-Wide Efficient Mixed Model Analysis:

The genome-wide efficient mixed model analysis (GEMMA) model is very similar to the FaST-LMM method (Zhou & Stephens, 2012). The GEMMA model also rewrites the likelihood function to one that is easier to evaluate and does not require that the variance parameters are the same across all SNPs (Zhou & Stephens, 2012; Tibbs Cortes et al., 2021). This model differs from FaST-LMM by using all markers to produce kinship results, which produces results like the EMMA method, only with improved speed (Zhou & Stephens, 2012; Tibbs Cortes et al., 2021).

Compressed Mixed Linear Model:

Different from the previously reviewed methods, the compressed mixed linear model (CMLM) has the goal of improving power (Zhang et al., 2010). This method aims

to do so by using a lower-rank kinship matrix (reviewed in Tibbs Cortes et al., 2021). CMLM uses unweighted pair-group method with arithmetic mean clustering and then uses the group means of the pair wise values to calculate kinship between groups (reviewed in Tibbs Cortes et al., 2021).

The enriched compressed mixed linear model (ECMLM) also uses a lower rank kinship matrix to improve power (Tibbs Cortes et al., 2021). Different from CMLM, ECMLM adds two additional parameters that need to be optimized. These parameters are the algorithm to cluster the individuals of the population into groups, and the method used to calculate the kinship between the groups (Li et al., 2014). This model then uses the P3D method to maximize the fit of the model before adding the marker effects (Li et al., 2014; Zhang et al., 2010). The ECMLM method provides the greater increase in power, when compared to CMLM method, but at a slower speed (Li et al., 2014).

Settlement of MLM Under Progressively Exclusive Relationship:

The settlement of MLM under progressively exclusive relationship (SUPER) method, created by Wang et al. (2014), also has the goal of calculating the kinship matrix with a higher speed. SUPER uses pseudo quantitative trait nucleotides (QTN) to create a reduced kinship matrix (Wang et al., 2014). The QTNs are calculated from when the SNPs are divided into bins, and the SNP with the lowest *p* value is designated as the pseudo QTN (Wang et al., 2014). SUPER is considered more powerful than FaST-LMM-Select, but it has lower computational efficiency (Tibbs Cortes et al., 2021).

Multi-Locus Mixed Model:

Multi-locus methods are used as a way to improve power of the calculations. The first multi-locus method was created by Segura et al. (2012) and named the multi-locus

mixed model (MLMM). Multi-locus models increase power by using multiple markers in the model concurrently as covariates (Tibbs Cortes et al., 2021). MLMM uses an iterative approach, where in each step, the genetic and error variance are estimated and used to calculate *p* values for the SNP and trait of interest (Segura et al., 2012). Then the most significant SNP is added to the model, and the process repeats until a threshold is met. After this process, a backward stepwise regression is used to remove the least significant SNP from the model until the optimal model is created (Segura et al., 2012).

Fixed and Random Model Circulating Probability Unification:

Another multi-locus model is the fixed and random model circulating probability unification (FarmCPU) method created by Liu et al. (2016). This method uses the reduced-rank kinship matrix of SUPER and iterates between the fixed effect model based on MLMM and the random-effect model of SUPER while taking into account restricted maximum likelihood (REML) as the optimization criteria (Liu et al., 2016; Tibbs Cortes et al., 2021). This method is increasingly popular and has been modified to work in multiple coding languages.

Bayesian Information and LD Iteratively Nested Keyway:

The FarmCPU method was then modified by its creators to develop the Bayesian information and LD iteratively nested keyway (BLINK) method (Huang et al., 2018). This method differs from FarmCPU by removing the requirement of SUPER that QTNs must be evenly distributed throughout the genome. This method recognizes that QTNs are often found in clusters throughout the genome (reviewed in Tibbs Cortes et al., 2021). The BLINK method also improves speed and efficiency with using a fixed effect model instead of the random effect model (Huang et al., 2018).

There are numerous GWAS methods to choose from, and a lot of the methods incorporate other models in them for higher power, optimization, or efficiency. GWAS and association mapping is an incredible powerful tool that many researchers exploit to provide insightful results within their research. With the popularity and continued advancement of GWAS research, there have been many tools and software packages created for user friendly options. Some of them include TASSEL (Bradbury et al., 2007), GAPIT (Lipka et al., 2012), and GEMMA (Zhou & Stephens, 2012).

Analysis of genetic control of phenolic compound accumulation in plants

Results from GWAS analysis can be used to make informed plant breeding decisions. GWAS can be used to help uncover genes that have potential effects in controlling the accumulation and production of specific phenolic compounds. Flavonoids and phenolic compounds play essential roles in most plants and crops and research delving into genetic control of phenolic compounds could lead to numerous advancements. In wheat, Chen et al. (2020), found 1098 significant SNP associations which lead to 26 candidate genes that could be a part of the control of a major flavonoid pathway. This study helped provide some of the initial steps in metabolomic-associated breeding of wheat (Chen et al., 2020). Another study, in barley, identified two markers as the major QTLs controlling phenolic compound content (Han et al., 2018). Additionally, Han et al. (2018) was able to identify a marker associated with the UDP-glycosyltransferase gene (*HvUGT*), which is a homolog to a gene in Arabidopsis that is confirmed to be involved in the biosynthesis of flavonoid glycosides. In mandarin, 420 SNPs were found in associated with 28 phenolic compounds in peel, pulp, or seed samples (Mattia et al., 2022). With the significant SNP results, four candidate genes were identified to be involved

in flavonoid biosynthesis, with a future goal of using the genes and markers to select mandarins with improved phenolic compound content for future breeding goals (Mattia et al., 2022). These are only a highlighted few of the multiple research studies looking to explore genetic control of phenolic compound biosynthesis and accumulation.

Flavonoids play an important role in maize. Previous research has been done to detect natural variation in flavonoid accumulation and biosynthesis in maize plants, as well as the functions carried out by these molecules (Jin et al., 2017; Wen et al., 2014). Research completed by Jin et al. (2017) identified 25 QTL corresponding to 23 different flavonoids across multiple maize populations. This research also uncovered 39 genes through an expression-based network analysis coupled with genetic mapping connected with flavonoid biosynthesis (Jin et al., 2017). Another study done in maize found 1,459 significant locus-trait association across multiple environments for metabolomics (Wen et al., 2014). This research then found 5 candidate genes involved with metabolomic traits, which could be used to influence future breeding decisions (Wen et al., 2014). More research on maize flavonoid biosynthesis and accumulation control in the kernels is needed to inform breeding decisions. Ability to control the content and accumulation of phenolic compounds in maize kernels can provide an increase in disease resistance, plant protection, and dietary benefits of consumption for both humans and animals.

PART 3 - FOURIER TRANSFORM INFRARED SPECTROSCOPY: MODELING OF PHENOLIC COMPOUND ACCUMULATION

History of plant phenotyping

Over the years, plant breeding has made numerous advancements in order to continue to feed our growing world. To continue to increase crop production and speed up the process of plant breeding, new techniques are needed to keep up with the growing demands. One of the

difficult bottlenecks of plant breeding is the time-consuming act of field phenotyping, especially when it is needed at multiple time points in multiple locations over multiple growing seasons (Kumar et al., 2015). The phenotype of a plant can be defined as the bases of morphological, biochemical, physiological, and molecular characteristics (Kumar et al., 2015). Plant phenotyping has been used as a means of variety selection for hundreds of years. It is the original method used for crop domestication (reviewed in Kumar et al., 2015).

Phenotyping has allowed for numerous advancements within plant breeding. One of the larger regions of study within phenotyping is phenotype plasticity of plants when exposed to stressors (Pieruschka & Schurr, 2019). Phenotyping has let researchers select varieties that are disease, drought, and pest resistant to keep up with changing and variable environments (Pieruschka & Schurr, 2019). Phenotyping has also allowed for the selection of high yielding and high nutritional varieties to keep up with the food demand of our growing world. Commonly, plant phenotyping is used with genetic data for genotype selection and to make informed decisions about crop varieties. This has allowed researchers to uncover the genetic control behind the important phenotypes of study and to make faster advancements in breeding programs.

Traditional phenotyping of crops is a labor intensive, time consuming, and high-cost practice that is essential for the furthering of plant breeding programs. This bottleneck in an essential part of the plant breeding process has created interest in creating more high-throughput methods of phenotyping. Techniques created to counteract the caveats of traditional phenotyping include noninvasive imaging, spectroscopy, image analysis, robotics, and high-performance computing (Kumar et al., 2015). The use of high throughput techniques are less invasive methods that can gather vast amounts of information on a larger scale. These developing

methods are faster than traditional phenotyping measures, have lower labor requirements, and reduce the bias of humans collecting data.

Phenomics

Phenomics stems from the word phenome, which refers to the phenotype as a whole, or on a larger scale (Kumar et al., 2015). Phenome provides large scale high-dimensional phenotypic data on the plant as a whole (Houle et al., 2010). Many of the problems plant breeders address are complex and without the use of large scale phenomic data, progression may be limited. Phenomic data acquisition can provide novel insight on how genomic variants are connected to phenotypes, which may help plant breeders understand more of their complex research interests (Houle et al., 2010). Some of the techniques included in phenomics research are infrared imaging, 3D imaging, magnetic resonance imaging, florescence imaging, and spectral reflectance (Pasala & Pandey, 2020). The use of spectral reflectance data is becoming increasingly important in plant research programs. Phenomic data acquisition techniques allow rapid advancements to be made in plant research such as understanding processes and mechanisms, rapid screening, forward and reverse genetic analysis, and production of elite plant varieties (Pasala & Pandey, 2020).

Spectroscopy

Spectroscopy is the study of the reflectance and absorption of light by matter. Spectroscopy is used as a technique that involves splitting light, or electromagnetic radiation, into a spectrum of wavelengths, in order to gain more knowledge on the properties of the matter. Spectroscopy and the use of spectral reflectance data is becoming an increasingly popular

technique for data collection; it is a precise and non-destructive technique that produces no harm to the plant or tissues being studied. By using spectroscopy as a means of data gathering, previously inaccessible plant phenotypes can be uncovered (Kalendar et al., 2022). Spectroscopy provides insight to properties of biomolecules and metabolites, biotic stress detection, abiotic stress detection, plant quality or health assessment, and identification of composition (reviewed in Kalendar et al., 2022). Spectral imaging techniques can also provide a means of early detection for stressors, before one would catch with the naked eye (Kalendar et al., 2022).

There are many different methods and techniques for gathering spectral data. The methods differ based on what is being analyzed, the type of interaction being monitored (absorption, emission, or diffraction), and the region of the electromagnetic spectrum used (Penner, 2017). In plant research, it is common to see methods based on the absorption or emission of radiation in the ultraviolet (UV), visible (Vis), infrared (IR), and radio or nuclear magnetic resonance (NMR) frequency ranges (Penner, 2017).

Spectroscopy provides a high-throughput technique for data gathering with a wide range of applications. Infrared and near-infrared spectroscopy is known for monitoring the reflection and absorbance of biomolecules in that range of wavelengths (Kalendar et al., 2022). NIR spectroscopy is the most common analytical technique applied in food quality research (Chandrasekaran et al., 2019). In food quality research, this technique has contributed to increased quality monitoring, composition analysis, and the sorting or grading of food items based on visual characteristics (Chandrasekaran et al., 2019). Success has also been found in monitoring key metabolites in grapevine organs on a large scale throughout the growing season (Wyngaard et al., 2021). Another impressive application of NIR spectroscopy is the determination of sorghum seed composition researched by Hacisalihoglu et al. (2022). Raman

spectroscopy (RS) is a method known for its effectiveness in analyzing the chemical structure of tissues (Cialla-May et al., 2022). As well as biochemistry applications, plant pathology, agronomy, and physiology applications have been evolving as a means of assessing overall plant health (reviewed in Kalendar et al., 2022).

Fourier transform infrared spectroscopy (FT-IR)

Fourier transform infrared (FT-IR) spectroscopy is a technique that uses infrared light to scan and observe the chemical properties of samples (Berna, 2017). FT-IR spectroscopy captures information on how IR light changes the dipole moments in molecules and how it responds to specific vibrational energy (JASCO inc., 2023). Historically, IR spectroscopy was a time-consuming process by individually checking the absorption of each frequency of IR light (Bruker, 2023). FT-IR spectroscopy has overtaken the historical techniques because it can check all of the wavelengths of IR light at the same time leading to a much faster process (Bruker, 2023). All chemical compounds create a unique spectral fingerprint that can be used for identification, allowing the results of FT-IR spectroscopy to create a unique profile of components for a sample (ThermoFisher Scientific, 2013). FT-IR has the ability to identify unknown materials, determine the quality of a sample, and can quantify the components in a mixture (Berna, 2017). FT-IR spectroscopy is an evolving tool that could provide huge advancements with quantification and detection of phenolic compounds in plant tissues.

Applications of FT-IR spectroscopy

FT-IR spectroscopy is a powerful technique that can be used in many different fields of study. One of the applications of FT-IR spectroscopy is the use in food quality and safety

research. Mohamed et al. (2011) used FT-IR spectroscopy on jams and juices that have been altered with synthetic flavors throughout the production process. This study found that there were specific peaks that correlated with synthetic pigments and flavors that were not found in natural juices (Mohamed et al., 2011). This research also concluded that FT-IR spectroscopy would be a powerful tool for detecting any adulterants in juices and jam adding to food safety research as well as the quality research completed (Mohamed et al., 2011). Another study on food quality used this technique to determine the contents and characteristics of flour (Sujka et al., 2017). This research found that FT-IR spectroscopy was a rapid and precise method that was successful in detection of content in flour ingredients (Sujka et al., 2017).

Another popular application of FT-IR spectroscopy is the use in soil sciences. This technique has been used for identifying and characterizing complex organic macromolecules found in soil (Stevenson, 1982). Research studies have been successfully done detecting the soil organic matter content in soils with different management practices using FT-IR spectroscopy (Ellerbrock et al., 2003). Another use of FT-IR spectroscopy in soils is the detection and analysis of microplastics (Park & Kim, 2022). This study looked at the size and accumulation in microplastic particles in different agricultural soil use types (Park & Kim, 2022). Park & Kim (2022) were able to use FT-IR to successfully identify the microplastic particles in the different soil types and compare sizes and amount.

FT-IR spectroscopy has many applications in a variety of disciplines, especially in research. Another discipline that frequently uses the identification properties of this tool is forensic science. FT-IR spectroscopy is a successful data collection method with analyzing textile synthetic fibers often involved as evidence in crime scenes (Aljannahi et al., 2022). This study was able to identify and group synthetic fibers that gives forensic science a powerful

technique with linking the fiber evidence to a suspect, victim and crime (Aljannahi et al., 2022). In addition to identification of synthetic fibers, FT-IR spectroscopy is also used on paint samples for forensic science purposes. Specifically with spray paint, Sharma et al. (2021) was able to conduct comparative studies between purposeful spray paint application and spray paint's overspray or contamination on other materials like gloves, shoes, wood, and hair. This research study also conducted a 100% accurate blind validation test linking spray paint samples with their origin (Sharma et al., 2021). Chemometric analysis capabilities of FT-IR provide huge advancement possibilities for the use of FT-IR in forensic science and criminal investigations.

Another application of FT-IR spectroscopy is in the nutrient analysis of plant tissues and products. Nutrient analysis normally takes analytical chemistry for identification and quantification. The use of FT-IR spectroscopy will reduce the time constraints previously introduced by the wet lab chemistry procedures. A study done by Bachhar et al. (2023) found success in identifying and confirming the presence of phenols, alkanes, aliphatic primary amines, carboxylic acids, nitrile, aromatics, and alcohols. The confirmed identification of these compounds led to information about the nutritional elements in the plant tissue including proteins, vitamins, carbohydrates, and amino acids (Bachhar et al., 2023). Another research study done by Mierzwa-Herszetek et al. (2019) looked at the total content of phenolic compounds using FT-IR analysis. This study used FT-IR spectroscopy to analyze the exogenous organic matter from plant biomass for the total phenolic content (Mierzwa-Herszetek et al., 2019). This study was able to successfully detect the total phenolic content of different biochars produced and to detect if there was any degradation of decomposition of molecules at different temperatures (Mierzwa-Herszetek et al., 2019). Using the knowledge learned from previous studies on FT-IR spectroscopy applications, it can be inferred that individual phenolic

compounds will have a unique spectral fingerprint that could be detected and quantified with this technique. A review done by Krysa et al. (2022) published the unique spectral bands for each group of flavonoids. This research will provide insight to future studies done analyzing individual flavonoid components using FT-IR spectroscopy (Krysa et al., 2022). Using FT-IR spectroscopy to detect phenolic compounds in ground maize kernels could provide new information about nutrient content, a rapid detection method for phenolic compounds, and validate time consuming analytical chemistry previously used for detection.

Using modeling techniques to understand data structure and convey results

Regression modeling is a group of techniques used to help researchers understand the data they collect and to make sense of the results. Regression modeling is commonly used for variable effect estimation and prediction purposes. Building and selecting the correct regression model depends on the structure of the data, sample size, possibility overfitting, and assessment of performance (Núñez et al., 2011). Both random forest and partial least squares are regression modeling techniques commonly used in research for effect estimation and prediction.

Random forest is classified as a statistical or machine learning algorithm, commonly used for prediction (Breiman, 2001; Schonlau & Zou, 2020). The random forest algorithm is created using tree-based models as building blocks (Schonlau & Zou, 2020). A tree-based model involves recursively partitioning the dataset into two groups based on a specific criterion; these tree-based models can either be based on classification tasks, for categorial outcomes, or regression tasks, for continuous outcome (Schonlau & Zou, 2020). For regression-based data, mean squared error is commonly used as a splitting criterion of the decision tree (Schonlau & Zou, 2020). Although powerful, decision trees and tree-based models can be predisposed to

overfitting. With the random forest algorithm, Breiman (2001) addresses the tendency of decision trees to be overfit by creating an algorithm that uses numerous individual trees, this increases the generalization accuracy (Schonlau & Zou, 2020). The random forest algorithm may be difficult to interpret, due to the numerous decision trees, but it often performs very well on prediction tasks (Breiman, 2001).

Partial least squares analysis is a statistical technique that allows researchers to compare multiple response and explanatory variables. This is done by combining features from principal component analysis and multiple linear regression (Abdi, 2003). The partial least squares technique addresses some of the drawbacks from using multiple linear regression by accounting for highly colinear variables, numerous factors and variables, and a relationship between the variables that is not well understood (Tobias, 1995). In regression problems, partial least squares assumes that most of the variation measured in the data is due to multiple underlying latent variables (Mehmood & Ahmed, 2015). The general underlying algorithm of partial least squares is that a set of explanatory variables are linked to response variables through a specific relationship with unknown regression parameters (Mehmood & Ahmed, 2015). The objective of the partial least squares technique is to predict the response by using the fitted model created by the regression parameters (Mehmood & Ahmed, 2015). This is done through the algorithm finding a set of components, or latent vectors, that explain the largest possible amount of covariance without overfitting, this is done similarly to a principal component analysis (Abdi, 2003). This modeling method is a multivariate technique proven to be effective in versatile fields including machine learning, bioinformatics, computer vision, agricultural research, and more (Mehmood & Ahmed, 2015).

While both random forest and partial least squares are successful and popular modeling

techniques, they both use different methods to estimate variable effects and make predictions. The differences in the methods behind the technique lead for different performances in different cases. A research study that looks at both techniques is completed by Lee et al. (2018), who used both random forest and partial least squares regression models to describe the relationship between phenolics and bioactivities of *Neptunia oleracea*. This study found that while both regression modeling techniques were useful and had strengths and weaknesses, random forest performed slightly poorer than partial least squares in prediction performance (Lee et al., 2018). A research study on the mapping of pasture biomass was found to prefer partial least squares regression over random forest, not for performance reasons, but for the lower computational power needed for the model (Otgonbayar et al., 2018). Another research study done on comparing both nonlinear and linear modeling methods for a near infrared calibration of paracetamol samples found that non-linear regression techniques were more successful than partial least squares regression (Sow et al., 2022). Contrasting the model decision from Lee et al. (2018) and Otgonbayar et al. (2018), Sow et al. (2022) had higher accuracy with random forest than partial least squares.

As seen in the research studies above, there is a wide range of data types and applications that regression modeling can be used for. Both random forest and partial least squares are highly powerful tools that are known for the ability to accurately estimate the variable effects and make predictions based on their regression. These techniques have very different algorithms behind the estimations and predictions, while random forest is a nonlinear approach that uses numerous decision trees, partial least square analysis is a multivariate linear based model that uses features of a principal component analysis (Breiman, 2001; Abdi, 2003). Future research that intends to employ regression modeling techniques should explore the use of both methods before making a

final decision. Regression model selection is a process based on data type and structure, computational efficiency, and performance. Exploring the use of both models with highly complex data for the best prediction accuracy may allow for a more informed selection process leading to the best results.

REFERENCES

Abdi, H. (2003). Partial least square regression (PLS regression). *Encyclopedia for research methods for the social sciences*, *6*(4), 792-795.

Aljannahi, A., Alblooshi, R. A., Alremeithi, R. H., Karamitsos, I., Ahli, N. A., Askar, A. M., Albastaki, I. M., Ahli, M. M., & Modak, S. (2022). Forensic analysis of textile synthetic fibers using a FT-IR spectroscopy approach. *Molecules*, *27*(13), 4281. https://doi.org/10.3390/molecules27134281

Babu, P. V., Liu, D., & Gilbert, E. R. (2013). Recent advances in understanding the anti-diabetic actions of dietary flavonoids. *The Journal of Nutritional Biochemistry*, *24*(11), 1777–1789. https://doi.org/10.1016/j.jnutbio.2013.06.003

Bachhar, V., Joshi, V., Gangal, A., Duseja, M., & Shukla, R. K. (2023). Identification of bioactive phytoconstituents, nutritional composition and antioxidant activity of calyptocarpus vialis. *Applied Biochemistry and Biotechnology*. https://doi.org/10.1007/s12010-023-04640-5

Badu-Apraku, B. & Fakorede, M. A. B. (2017). Molecular Approaches to Maize Improvement. In: Advances in Genetic Enhancement of Early and Extra-Early Maize for Sub-Saharan Africa. Springer, Cham. https://doi.org/10.1007/978-3-319-64852-1_8

Bais, H. P., Vepachedu, R., Gilroy, S., Callaway, R. M., & Vivanco, J. M. (2003). Allelopathy and exotic plant invasion: From molecules and genes to species interactions. *Science, 301*(5638), 1377-1380. https://doi.org/10.1126/science.1083245

Bajet, N.B., Renfro, B.L., & Carrasco, J.M.V. (1994). Control of tar spot of maize and its effect on yield. *Int. J. Pest Manage*. 40:121-125. https://doi.org/10.1080/09670879409371868

Ballard, C. R. & Junior, M. R. M. (2019). Chapter 10 - Health Benefits of Flavonoids. In M. R. S. Campos (Ed.), *Bioactive Compounds* (pp. 185–201). essay, Woodhead Publishing. https://doi.org/10.1016/B978-0-12-814774-0.00010-4

Bataglion, G. A., da Silva, F. M. A., Eberlin, M. N., & Koolen, H. H. F. (2015). Determination of the phenolic composition from Brazilian tropical fruits by UHPLC–ms/MS. *Food Chemistry*, *180*, 280–287. https://doi.org/10.1016/j.foodchem.2015.02.059

Bashandy, T., Taconnat, L., Renou, J. P., Meyer, Y., & Reichheld, J. P. (2009). Accumulation of flavonoids in an NTRA ntrb mutant leads to tolerance to UV-C. *Molecular Plant*, *2*(2), 249–258. https://doi.org/10.1093/mp/ssn065

Basli, A., Belkacem, N., & Amrani, I. (2017). Health benefits of phenolic compounds against cancers. *Phenolic Compounds - Biological Activity*. https://doi.org/10.5772/67232

Bennett, R. J. & Turgeon, B. G. (2016). Fungal sex: The *Ascomycota*. *Microbiology Spectrum*, *4*(5). https://doi.org/10.1128/microbiolspec.funk-0005-2016

Benson, J. M., Poland, J. A., Benson, B. M., Stromberg, E. L., & Nelson, R. J. (2015). Resistance to gray leaf spot of maize: Genetic architecture and mechanisms elucidated through nested association mapping and near-isogenic line analysis. *PLoS Genetics*, **11**, e1005045. https://doi.org/10.1371/journal.pgen.1005045

Berna, F. (2017). Fourier Transform Infrared Spectroscopy (FTIR). In: Gilbert, A.S. (eds) Encyclopedia of Geoarchaeology. Encyclopedia of Earth Sciences Series. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4409-0_15

Bernardo, R. (2020). *Breeding for quantitative traits in plants* (Third edition). Stemma Press.

Bowles D., Isayenkova J., Lim E. K., Poppenberger B. (2005). Glycosyltransferases: managers of small molecules. *Current Opinion in Plant Biology*. *8*(3), 254–263. https://doi.org/10.1016/j.pbi.2005.03.007

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). Tassel: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Bredsdorff, L., Nielsen, I. L., Rasmussen, S. E., Cornett, C., Barron, D., Bouisset, F., Offord, E., & Williamson, G. (2010). Absorption, conjugation and excretion of the flavanones, naringenin and hesperetin from α-rhamnosidase-treated orange juice in human subjects. *British Journal of Nutrition*, *103*(11), 1602–1609. https://doi.org/10.1017/s0007114509993679

Breiman L. (2001). Random forests. *Machine Learning* 45: 5–32

Bruker. (2023). *Guide to FT-IR spectroscopy*. Spectroscopy Basics. https://www.bruker.com/en/products-and-solutions/infrared-and-raman/ft-ir-routine-spectrometer/what-is-ft-ir-spectroscopy.html

Cao, S., Loladze, A., Yuan, Y., Wu, Y., Zhang, A., Chen, J., Huestis, G., Cao, J., Chaikam, V., Olsen, M., Prasanna, B. M., San Vicente, F., & Zhang, X. (2017). Genome-wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction. *The Plant Genome*, *10*(2). https://doi.org/10.3835/plantgenome2016.10.0099

Cannon, P. F. (1991). A revision of *Phyllachora* and some similar genera on the host family Leguminosae. *Mycological Papers*, 163-302.

Cano-Gamez, E. & Trynka, G. (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00424

Chandrasekaran, I., Panigrahi, S. S., Ravikanth, L., & Singh, C. B. (2019). Potential of near-infrared (NIR) spectroscopy and hyperspectral imaging for quality and safety assessment of fruits: An overview. *Food Analytical Methods*, *12*(11), 2438–2458. https://doi.org/10.1007/s12161-019-01609-1

Chaves, J. O., Corrêa de Souza, M., Capelasso da Silva, L., Lachos-Perez, D., Torres-Mayanga, P. C., Paula da Fonseca Machado, A., Forster-Carneiro, T., Vázquez-Espinosa, M., González-de-Peredo, A. V., Barbero, G. F., & Rostagno, M. A. (2020). Extractions of Flavonoids From Natural Sources Using Modern Techniques. *Frontiers in Chemistry: Green and Sustainable Chemistry, 8.* https://doi.org/10.3389/fchem.2020.507887

Chen, J., Hu, X., Shi, T., Yin, H., Sun, D., Hao, Y., Xia, X., Luo, J., Fernie, A. R., He, Z., & Chen, W. (2020). Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. *Plant Biotechnology Journal*, *18*(8), 1722–1735. https://doi.org/10.1111/pbi.13335

Cialla-May, D., Krafft, C., Rösch, P., Deckert-Gaudig, T., Frosch, T., Jahn, I. J., Pahlow, S., Stiebing, C., Meyer-Zedler, T., Bocklitz, T., Schie, I., Deckert, V., & Popp, J. (2021). Raman spectroscopy and imaging in Bioanalytics. *Analytical Chemistry*, *94*(1), 86–119. https://doi.org/10.1021/acs.analchem.1c03235

Cline, E. (2005). Phyllachora maydis. U.S. National Fungus Collections, ARS, USDA. Retrieved June 30, 2021, from https://nt.ars-grin.gov/sbmlweb/fungi/nomenSheets.cfm

Cocuron, J. C., Casas, M. I., Yang, F., Grotewold, E., & Alonso, A. P. (2019). Beyond the wall: High-throughput quantification of plant soluble and cell-wall bound phenolics by liquid chromatography tandem mass spectrometry. *Journal of Chromatography A*, *1589*, 93-104. https://doi.org/10.1016/j.chroma.2018.12.059

Corn ipmPIPE. (2023, February 23). Historical end of season maps. Tar Spot. https://corn.ipmpipe.org/tarspot/historical-end-of-season-maps/

Crop Protection Network (2023). Estimates of crop yield losses due to diseases and invertebrate pests: an online tool https://loss.cropprotectionnetwork.org/ https://doi.org/10.31274/cpn-20191121-0 (Accessed July 18, 2023).

Dell'Acqua, M., Gatti, D. M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., Hlaing, A. L., Aung, H. H., Neilissen, H., Baute, J., Frascaroli, E., Churchill, G., Inzé, D., Morgante, Michele, & Pè, M. E. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biology, 16*(167). https://doi.org/10.1186/s13059-015-0716-z

Dudareva, N., Pichersky, E., & Gershenzon, J. (2004). Biochemistry of Plant Volatiles. *Plant Physiology, 135*(4), 1893-1902. https://doi.org/10.1104/pp.104.049981

Ellerbrock, R. H., Höhn, A., & Rogasik, J. (2003). Functional analysis of soil organic matter as

affected by long-term manurial treatment. *European Journal of Soil Science*, *50*(1), 65–71. https://doi.org/10.1046/j.1365-2389.1999.00206.x

Erlund, I. (2004). Review of the flavonoids quercetin, hespertin, and naringenin. Dietary sources, bioactivities, bioavailability, and epidemiology. *Nutritional Resources, 24*, 851-874. https://doi.org/10.1016/j.nutres.2004.07.005

Falcone Ferreyra, M. L., Rius, S. P., & Casati, P. (2012). Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Frontiers in plant science*, *3*, 222. https://doi.org/10.3389/fpls.2012.00222

Falconer, D. S. & Mackay, T. F. C. (1996) Introduction to Quantitative Genetics. 4th Edition, Addison Wesley Longman, Harlow.

Ferrer J., Austin M., Stewart C. J., & Noel J. (2008). Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. *Plant Physiology and Biochemistry*. *46*(3), 356–370. https://doi.org/10.1016/j.plaphy.2007.12.009

FTIR Spectroscopy Basics: Thermo Fisher Scientific - US. FTIR Spectroscopy Basics | Thermo Fisher Scientific - US. (2013). https://www.thermofisher.com/us/en/home/industrial/spectroscopy-elemental-isotope-analysis/spectroscopy-elemental-isotope-analysis-learning-center/molecular-spectroscopy-information/ftir-information/ftir-basics.html.

Groves, G. L., Kleczewski, N. M., Telenko, D. E. P., Chilvers, M. I., & Smith, D. L. (2020). *Phyllachora maydis* ascospore release and germination from overwintered corn residue. *Plant Health Progress*, *21*,1. https://doi.org/10.1094/PHP-10-19-0077-RS

Gruz, J., Novák, O., & Strnad, M. (2008). Rapid analysis of phenolic acids in beverages by UPLC–ms/MS. *Food Chemistry*, *111*(3), 789–794. https://doi.org/10.1016/j.foodchem.2008.05.014

Habtemariam, S. & Varghese, G. (2014). The antidiabetic therapeutic potential of dietary polyphenols. *Current Pharmaceutical Biotechnology*, *15*(4), 391–400. https://doi.org/10.2174/1389201015666140617104643

Hacisalihoglu, G., Armstrong, P. R., Mendoza, P. T., & Seabourn, B. W. (2022). Compositional analysis in sorghum (*Sorghum bicolor*) nir spectral techniques based on mean spectra from single seeds. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.995328

Han, Z., Zhang, J., Cai, S., Chen, X., Quan, X., & Zhang, G. (2018). Association mapping for total polyphenol content, total flavonoid content and antioxidant activity in Barley. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-4483-6

Hock, J., Kranz, J., & Renfro, B. (1989). El complejo ‘mancha de asfalto’ de maíz: Su

distribucción geográfica, requisitos ambientales e importancia económica en México. *Revista Mexicana de Fitopatologia*, 7(2), 129–35.

Hock, J., Dittrich, U., Renfro, B. L., & Kranz, J. (1992). Sequential development of pathogens in the maize tarspot disease complex. *Mycopathologia*, 117(3): 157–161. https://doi.org/10.1007/BF00442777

Hock, J., Kranz, J., & Renfro B. L. (1995). Studies on the epidemiology of the tar spot disease complex of maize in Mexico. *Plant Pathology*, 44(3):490–502. https://doi.org/10.1111/j.1365-3059.1995.tb01671.x

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: The next challenge. *Nature Reviews Genetics*, *11*(12), 855–866. https://doi.org/10.1038/nrg2897

Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2018). Blink: A package for the next level of genome-wide association studies with both individuals and markers in the Millions. *GigaScience*, *8*(2). https://doi.org/10.1093/gigascience/giy154

JASCO Inc. (2023, May 17). *FTIR spectroscopy (overview)*. Fundamental Theory and Applications of FTIR Spectroscopy. https://jascoinc.com/learning-center/theory/spectroscopy/fundamentals-ftir-spectroscopy/#FTIR-spectroscopy-principles

Jin, M., Zhang, X., Zhao, M., Deng, M., Du, Y., Zhou, Y., Wang, S., Tohge, T., Fernie, A.R., Willmitzer, L., Brotman, Y., Yan, J., & Wen, W. (2017). Integrated genomics-based mapping reveals the genetics underlying maize flavonoid biosynthesis. *BMC Plant Biology, 17*, 17. https://doi.org/10.1186/s12870-017-0972-z

Kalendar, R., Ghamkhar, K., Franceschi, P., & Egea-Cortines, M. (2022). Editorial: Spectroscopy for crop and product phenotyping. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.1058333

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in Model Organism Association mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kao, C. H., Zeng, Z. B., & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics, 152*(3), 1203-1216. https://doi.org/10.1093/genetics/152.3.1203

Krysa, M., Szymańska-Chargot, M., & Zdunek, A. (2022). FT-IR and FT-Raman fingerprints of flavonoids – a review. *Food Chemistry*, *393*, 133430.

https://doi.org/10.1016/j.foodchem.2022.133430

Kumar, J., Pratap, A., & Kumar, S. (2015). Plant Phenomics: An Overview. In: Kumar, J., Pratap, A., Kumar, S. (eds) Phenomics in Crop Plants: Trends, Options and Limitations. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2226-2_1

Kozlowska, A., & Szostak-Węgierek, D. (2014). Flavonoids – Food sources and health benefits. *Yearbooks of the National Institute of Hygiene (Roczniki Państwowego Zakładu Higieny*), *65*(2), 79-85.

Lattanzio, V., Cardinali, A., Linsalata, V. (2012). Plant phenolics: a biochemical and physiological perspective. In: Cheynier, V., Sarni-Manchado, P., Quideau, S. (Eds.), *Recent Advances in Polyphenol Research*. John Wiley & Sons, Ltd, pp. 1–39. https://doi.org/10.1002/9781118299753.ch1

Lee, S. Y., Mediani, A., Maulidiani, M., Khatib, A., Ismail, I. S., Zawawi, N., & Abas, F. (2018). Comparison of partial least squares and random forests for evaluating relationship between phenolics and bioactivities of Neptunia oleracea. *Journal of the science of food and agriculture*, *98*(1), 240–252. https://doi.org/10.1002/jsfa.8462

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y. M., Todhunter, R. J., Buckler, E. S., & Zhang, Z. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biology*, *12*(1). https://doi.org/10.1186/s12915-014-0073-5

Lin, J., Rexrode, K. M., Hu, F., Albert, C. M., Chae, C. U., Rimm, E. B., Stampfer, M. J., & Manson, J. E. (2007). Dietary intakes of flavonols and flavones and coronary heart disease in US women. *American Journal of Epidemiology*, *165*(11), 1305–1313. https://doi.org/10.1093/aje/kwm016

Lin, Y., Xu, W., Huang, M., Xu, W., Li, H., Ye, M., Zhang, X., & Chu, K. (2015). Qualitative and quantitative analysis of phenolic acids, flavonoids and iridoid glycosides in Yinhua Kanggan tablet by UPLC-QQQ-MS/MS. *Molecules*, *20*(7), 12209–12228. https://doi.org/10.3390/molecules200712209

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., & Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics (Oxford, England)*, *28*(18), 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835. https://doi.org/10.1038/nmeth.1681

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., & Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. Nature Methods, 9(6), 525–526. https://doi.org/10.1038/nmeth.2037

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics, 12*(2). https://doi.org/10.1371/journal.pgen.1005767

Liu, L. J. (1973). Incidence of tar spot disease of corn in Puerto Rico. *J. Agric. Univ. Puerto Rico*, *42*, 211-216.

Liu, Z., Liu, Y., Pu, Z., Wang, J., Zheng, Y., Li, Y, & Wei, Y. (2013). Regulation, evolution, and functionality of flavonoids in cereal crops. *Biotechnology Letters 35*, 1765–1780. https://doi.org/10.1007/s10529-013-1277-4

Loladze, A., Rodrigues, F.A., Toledo, F., San Vicente, F., Gérard, B., & Boddupalli, M.P. (2019). Application of remote sensing for phenotyping tar spot complex resistance in maize. *Frontiers in Plant Science*, **10**, 552. https://doi.org/10.3389/fpls.2019.00552

Mahuku, G., Chen, J., Shrestha, R., Narro, L., Guerrero, K. V. O., Arcos, A. L., and Xu, Y. (2016). Combined linkage and association mapping identifies a major QTL (*qRtsc8-1*), conferring tar spot complex resistance in maize. *Theoretical and Applied Genetics, 129*, 1217-1229. https://doi.org/10.1007/s00122-016-2698-y

Mammadov, J., Sun, X., Gao, Y., Ochsenfeld, C., Bakker, E., Ren, R., Flora, J., Wang, X., Kumpatla, S., Meyer, D., & Thompson, S. (2015). Combining powers of linkage and association mapping for precise dissection of QTL controlling resistance to gray leaf spot disease in maize (*Zea mays* L.). *BMC Genomics*, **16**, 916. https://doi.org/10.1186/s12864-015-2171-3

Martens S., Preuss A., & Matern, U. (2010). Multifunctional flavonoid dioxygenases: flavonols and anthocyanin biosynthesis in Arabidopsis thaliana L. Phytochemistry 71, 1040–1049. https://doi.org/10.1016/j.phytochem.2010.04.016

Matthies, A., Clavel, T., Gütschow, M., Engst, W., Haller, D., Blaut, M., & Braune, A. (2008). Conversion of daidzein and Genistein by an anaerobic bacterium newly isolated from the mouse intestine. *Applied and Environmental Microbiology*, *74*(15), 4847–4852. https://doi.org/10.1128/aem.00555-08

Mattia, M. R., Du, D., Yu, Q., Kahn, T., Roose, M., Hiraoka, Y., Wang, Y., Munoz, P., & Gmitter, F. G. (2022). Genome-wide association study of healthful flavonoids among diverse Mandarin accessions. *Plants*, *11*(3), 317. https://doi.org/10.3390/plants11030317

Mehmood, T. & Ahmed, B. (2015). The diversity in the applications of partial least squares: An overview. *Journal of Chemometrics*, *30*(1), 4–17. https://doi.org/10.1002/cem.2762

Michel, K. J., Lima, D. C., Hundley, H., Singan, V., Yoshinaga, Y., Daum, C., Barry, K., Broman, K. W., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2022). Genetic mapping and prediction of flowering time and plant height in a maize Stiff Stalk MAGIC population. *Genetics, 221*(2). https://doi.org/10.1093/genetics/iyac063

Mierzwa-Hersztek, M., Gondek, K., Mierzwa-Hersztek, M., Nawrocka, A., Pińkowska, H., Bajda, T., Stanek-Tarkowska, J., & Szostek, M. (2019). FT-IR analysis and the content of phenolic compounds in exogenous organic matter produced from plant biomass. *Journal of Elementology*, (3/2019). https://doi.org/10.5601/jelem.2018.23.3.1716

Miles, C. & Wayne, M. (2008) Quantitative trait locus (QTL) analysis. *Nature Education, 1*(1):208

Mol, J., Grotewold, E., & Koes, R. (1998). How genes paint flowers and seeds. *Trends in Plant Science*, *3*(6), 212-217. https://doi.org/10.1016/S1360-1385(98)01242-4

Mohamed, G. F., Shaheen, M. S., Khalil, S. K., Hussein, A. M. S., & Kamil, M. M. (2011). Application of FT-IR spectroscopy for rapid and simultaneous quality determination of some fruit products. *Nat. Sci*, *9*(11), 21-31.

Mueller, D. S., Wise, K. A., Sisson, A. J., Allen, T. W., Bergstrom, G. C., Bissonnette, K. M., Bradley, C. A., Byamukama, E., Chilvers, M. I., Collins, A. A., Esker, P. D., Faske, T. R., Friskop, A. J., Hagan, A. K., Heiniger, R. W., Hollier, C. A., Isakeit, T., Jackson-Ziems, T. A., Jardine, D. J., & Wiebold, W. J. (2020). Corn yield loss estimates due to diseases in the United States and Ontario, Canada, from 2016 to 2019. *Plant Health Progress*, *21*(4), 238–247. https://doi.org/10.1094/php-05-20-0038-rs

Naranjo-Ortiz, M. A. & Gabaldón, T. (2019). Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biological reviews of the Cambridge Philosophical Society*, *94*(6), 2101–2137. https://doi.org/10.1111/brv.12550

National Institute of Health. (2020). *Genome-wide association studies fact sheet*. National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

Núñez, E., Steyerberg, E. W., & Núñez, J. (2011). Regression modeling strategies. *Revista Española de Cardiología (English Edition)*, *64*(6), 501–507. https://doi.org/10.1016/j.rec.2011.01.017

Olukolu, B. A., Tracy, W. F., Wisser, R., De Vries, B., & Balint-Kurti, P. J. (2016). A genome-wide association study for partial resistance to maize common rust. *Phytopathology, 106,* 745-751. https://doi.org/10.1094/PHYTO-11-15-0305-R

Otgonbayar, M., Atzberger, C., Chambers, J., & Damdinsuren, A. (2018). Mapping pasture biomass in Mongolia using partial least squares, random forest regression and Landsat 8 imagery. *International Journal of Remote Sensing*, *40*(8), 3204–3226. https://doi.org/10.1080/01431161.2018.1541110

Panche A.N., Diwan A.D., & Chandra S.R. (2016). Flavonoids: an overview. *Journal of Nutritional Science, 5*(47). https://doi.org/10.1017/jns.2016.41

Park, S. Y. & Kim, C. G. (2022). A comparative study on the distribution behavior of microplastics through FT-IR analysis on different land uses in agricultural soils. *Environmental Research*, *215*, 114404. https://doi.org/10.1016/j.envres.2022.114404

Pasala, R. & Pandey, B. B. (2020). Plant Phenomics: High-throughput technology for accelerating genomics. *Journal of Biosciences*, *45*(1). https://doi.org/10.1007/s12038-020-00083-w

Penner, M. H. (2017). Basic Principles of Spectroscopy. In: Nielsen, S.S. (eds) Food Analysis. Food Science Text Series. Springer, Cham. https://doi.org/10.1007/978-3-319-45776-5_6

Perez-Vizcaino, F. & Duarte, J. (2010). Flavonols and cardiovascular disease. *Molecular Aspects of Medicine*, *31*(6), 478–494. https://doi.org/10.1016/j.mam.2010.09.002

Pieruschka, R. & Schurr, U. (2019). Plant phenotyping: Past, present, and future. *Plant Phenomics*, *2019*. https://doi.org/10.34133/2019/7507131

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Roff, D. A. (2007). A centennial celebration for quantitative genetics. *Evolution; international journal of organic evolution*, *61*(5), 1017–1032. https://doi.org/10.1111/j.1558-5646.2007.00100.x

de la Rosa, L. A., Moreno-Escamilla, J. O., Rodrigo-Garcia, J., & Alvarez-Parrilla, E. (2018). Phenolic Compounds. In *Postharvest Physiology and Biochemistry of Fruits and Vegetables* (pp. 253–272). essay, Woodhead Publishing.

Ruhl, G., Romberg, M. K., Bissonnette, S., Plewa, D., Creswell, T., & Wise, K. A. (2016). First report of tar spot on corn caused by *Phyllachora maydis* in the United States. *Plant Disease, 100*(7), 1496. https://doi.org/10.1094/pdis-12-15-1506-pdn

Šamec, D., Karalija, E., Šola, I., Bok, V. V., & Salopek-Sondi, B. (2021). The role of polyphenols in abiotic stress response: The influence of molecular structure. *Plants, 10*(1)*,* 118. https://doi.org/10.3390/plants10010118

Schonlau, M., & Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867x20909688

Scott, M. F., Ladejobi, O., Amer, S., Bently, A. R., Biernaskie, J., Boden, S. A., Clark, M., Dell'Acqua, M., Dixon, L. E., Filippi, C. V., Fradgley, N., Gardner, K. A., Mackay, I. J.,

O'Sullivan, D., Percival-Alwyn, L., Roorkiwal, M., Singh, R. K., Thudi, M., Varshney, R. K., Venturini, L., Whan, A., Cockram, J., & Mott, R. (2020). Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity, 125*:396-416. https://doi.org/10.1038/s41437-020-0336-6

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, *44*(7), 825–830. https://doi.org/10.1038/ng.2314

Shahidi, F. & Naczk, M. (2003). Phenolics in Food and Nutraceuticals (Second Edition). CRC Press. https://doi.org/10.1201/9780203508732

Sharma, S., Chophi, R., Kaur, C., & Singh, R. (2021). Chemometric analysis on atr-ft-ir spectra of spray paint samples for forensic purposes. *Journal of Forensic Sciences*, *66*(6), 2190–2200. https://doi.org/10.1111/1556-4029.14806

Shi, L., Lu, X., Weng, J., Zhu, H., Liu, C., Hao, Z., Zhou, Y., Zhang, D., Li, M., Ci, X., Li, X., & Zhang, S. (2014). Genetic characterization and linkage disequilibrium mapping of resistance to gray leaf spot in maize (Zea mays L.). *The Crop Journal, 2,* 132-143. https://doi.org/10.1016/j.cj.2014.02.001

Soto-Vaca, A., Losso, J. N., Xu, Z., & Finley, J. W. (2021). Review: Evolution of phenolic compounds from color and flavor problems to health benefits. *Journal of Agriculture and Food Chemistry, 60*(27), 6658-6677. https://doi.org/10.1021/jf300861c

Sow, A., Traore, I., Diallo, T., Traore, M., & Ba, A. (2022). Comparison of gaussian process regression, partial least squares, random forest and support vector machines for a near infrared calibration of paracetamol samples. *Results in Chemistry*, *4*, 100508. https://doi.org/10.1016/j.rechem.2022.100508

Stevenson, F.J. 1982 *Humus Chemistry: Genesis, Composition, Reactions*. John Wiley & Sons, New York.

Sujka, K., Koczoń, P., Ceglińska, A., Reder, M., & Ciemniewska-Żytkiewicz, H. (2017). The application of FT-IR spectroscopy for quality control of flours obtained from Polish producers. *Journal of Analytical Methods in Chemistry*, *2017*, 1–9. https://doi.org/10.1155/2017/4315678

Swanson, B.G. (2003). Tannins and Polyphenols. *Encyclopedia of Food Sciences and Nutrition (Second Edition) Academic Press.* 5729-5733. https://doi.org/10.1016/B0-12-227055-X/01178-0

Telenko, D., Chilvers, M. I., Kleczewski, N., Mueller, D., Plewa, D., Robertson, A., Smith, D., Tenuta, A., & Wise, K. (2020). An overview of tar spot. *Crop Protection Network*. https://doi.org/10.31274/cpn-20190620-008

Telenko, D. E. P., Chilvers, M. I., Kleczewski, N., Smith, D. L., Byrne, A. M., Devillez, P., Diallo, T., Higgins, R., Joos, D., Kohn, K., Lauer, J., Mueller, B., Singh, M. P., Widdicombe, W. D., & Williams, L. A. (2019). How tar spot of corn impacted hybrid yields during the 2018 midwest epidemic. *Crop Protection Network*. https://doi.org/10.31274/cpn-20190729-002

Telenko, D. E. P., Chilvers, M. I., Kleczewski, N., Smith, D. L., Byrne, A. M., Devillez, P., Diallo, T., Higgins, R., Joos, D., Kohn, K., Lauer, J., Mueller, B., Singh, M. P., Widdicombe, W. D., & Williams, L. A. (2019). *How Tar Spot of Corn Impacted Hybrid Yields during the 2018 Midwest Epidemic*. https://cropprotectionnetwork.org/publications/how-tar-spot-of-corn-impacted-hybrid-yields-during-the-2018-midwest-epidemic https://doi.org/10.31274/cpn-20190729-002

Tibbs Cortes, L., Zhang, Z., & Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *The Plant Genome*, *14*(1). https://doi.org/10.1002/tpg2.20077

Tobias, R. D. (1995, April). An introduction to partial least squares regression. In *Proceedings of the twentieth annual SAS users group international conference* (Vol. 20, pp. 1250-1257). Cary, NC, USA: SAS Institute Inc.

Trygestad, B. (2021). Genetic And Genetic by Environment Effects on Tar Spot Resistance and Hybrid Yield in Maize (thesis).

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-021-00056-9

Valle-Torres, J., Ross, T. J., Plewa, D., Avellaneda, M. C., Check, J., Chilvers, M. I., Cruz, A. P., Dalla Lana, F., Groves, C., Gongora-Canul, C., Henriquez-Dole, L., Jamann, T., Kleczewski, N., Lipps, S., Malvick, D., McCoy, A. G., Mueller, D. S., Paul, P. A., Puerto, C., Schloemer, C., Raid, R. N., Robertson, A., Roggenkamp, E. M., Smith, D. L., Telenko, D. E. P., & Cruz, C. D. (2020). Tar Spot: An Understudied Disease Threatening Corn Production in the Americas. *Plant Disease 104*(10). https://doi.org/10.1094/PDIS-02-20-0449-FE

van Wyngaard, E., Blancquaert, E., Nieuwoudt, H., & Aleixandre-Tudo, J. L. (2021). Infrared spectroscopy and chemometric applications for the qualitative and Quantitative Investigation of Grapevine Organs. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.723247

Wang, L., Lee, I. M., Zhang, S. M., Blumberg, J. B., Buring, J. E., & Sesso, H. D. (2009). Dietary intake of selected flavonols, flavones, and flavonoid-rich foods and risk of cancer in middle-aged and older women. *The American Journal of Clinical Nutrition*, *89*(3), 905–912. https://doi.org/10.3945/ajcn.2008.26913

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., & Zhang, Z. (2014). A super powerful method for Genome Wide Association study. *PLoS ONE*, *9*(9). https://doi.org/10.1371/journal.pone.0107684

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W., Luo, J., & Yan, J. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nature Communications.* 5, 3438. https://doi.org/10.1038/ncomms4438

White, M.R., Mikel, M.A., Leon, N., & Kaeppler, S. M. 2020. Diversity and heterotic patterns in North American proprietary dent maize germplasm. *Crop Science, 60*(1):100–114. https://doi.org/10.1002/csc2.20050

Xiao, Z. P., Peng, Z. Y., Peng, M. J., Yan, W.-B., Ouyang, Y.-Z., & Zhu, H.-L. (2011). Flavonoids health benefits and their molecular mechanism. *Mini-Reviews in Medicinal Chemistry*, *11*(2), 169–177. https://doi.org/10.2174/138955711794519546

Xu, D. P., Li, Y., Meng, X., Zhou, T., Zhou, Y., Zheng, J., Zhang, J. J., & Li, H. B. (2017). Natural Antioxidants in Foods and Medicinal Plants: Extraction, Assessment and Resources. *International Journal of Molecular Sciences. 18*(1), 96. https://doi.org/10.3390/ijms18010096

Yan, S., Loladze, A., Wang, N., Sun, S., Chilvers, M. I., Olsen, M., Burgueño, J., Petroli, C. D., Molnar, T., San Vicente, F., Zhang, X., & Prasanna Boddupalli, M. (2022). Association mapping of resistance to tar spot complex in maize. *Plant Breeding*, *141*(6), 745–755. https://doi.org/10.1111/pbr.13056

Yu, J., & Buckler, E. S. (2006). Genetic association mapping and Genome Organization of Maize. *Current Opinion in Biotechnology*, *17*(2), 155–160. https://doi.org/10.1016/j.copbio.2006.02.003

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Yu, K., Little, D., Plumb, R., & Smith, B. (2006). High-throughput quantification for a drug mixture in rat plasma – a comparison of Ultra PerformanceTM Liquid Chromatography/tandem mass spectrometry with high-performance liquid chromatography/tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, *20*(4), 544–552. https://doi.org/10.1002/rcm.2336

Zeng, Z. B. (2001). QTL Mapping. *Brenner's Encyclopedia of Genetics*, 8–12. https://doi.org/10.1016/b978-0-12-374984-0.01248-1

Zeng, Z. B., Kao, C., & Basten, C. (1999). Estimating the genetic architecture of quantitative

traits. Genetics Research, 74(3), 279-289. https://doi/org/10.1017/S0016672399004255

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360. https://doi.org/10.1038/ng.546

Zhang, Z., Liang, Z., Yin, L., Li, Q. X., & Wu, Z. (2018). Distribution of Four Bioactive Flavonoids in Maize Tissues of Five Varieties and Correlation with Expression of the Biosynthetic Genes. *Journal of Agriculture and Food Chemistry, 66*(40), 10431-10437. https://doi.org/10.1021/acs.jafc.8b03865

Zheng, H., Chen, J., Mu, C., Makumbi, D., Xu, Y., & Mahuku, G. (2018). Combined linkage and association mapping reveal QTL for host plant resistance to common rust (*Puccinia sorghi*) in tropical maize. *BMC Plant Biology*, *18*, 310. https://doi.org/10.1186/s12870-018-1520-1

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

CHAPTER 2: VALIDATING GENETIC RESISTANCE TO MAIZE TAR SPOT IN A STIFF-STALK MAGIC POPULATION

ABSTRACT

Tar spot of maize is caused by the fungal pathogen *Phyllachora maydis*. Symptoms include the development of small black lesions on the foliar tissue of maize plants, resembling bits of tar. In-season management of tar spot is challenging, as once it is established it can spread rapidly in conducive environments and cause significant damage, reducing stalk integrity, grain yield, and forage quality. The most cost-effective management option in the long term will be to plant resistant hybrids; for this, alleles from resistant genotypes must be identified and incorporated into breeding programs. In this study, a stiff-stalk multiparent advanced-generation intercross (MAGIC) population of 500 maize lines was assessed in three different trial locations - Indiana, Michigan, and Wisconsin - to validate previously observed marker-trait associations and enable identification of new quantitative trait loci (QTL) in the parental genotypes. Stromatal severity ratings on the ear leaves of this population in the different locations were coupled with the marker data for the population and sequence data for the founders to conduct QTL analysis to validate and discover new resistant loci conferring tar spot disease resistance. Results of the QTL analysis led to the discovery of a major QTL on chromosome 1, at the position of 1.91 Mb from multiple of the traits and locations. Founders B73 and B84 confer the largest negative QTL effects, corresponding to resistance in the population. Future work includes screening of two doubled haploid populations with resistant lines crossed to either B73 or LH244. Identifying markers linked to validated QTLs for tar spot resistance will enable plant breeders to leverage these discoveries, breed for resistant varieties, and minimize the devastating impact of this fungal disease.

INTRODUCTION

Maize is one of the most widely grown crops, used for food, forage, and production materials. It is considered a staple crop depended on by most countries in the world. Constant pressure from pests, diseases, and environmental stressors threatens the stability of maize and the societies that depend on it. One of the new diseases causing detrimental effects to this staple crop is tar spot (Valle-Torres et al., 2020. Tar spot of maize is a foliar disease caused by the pathogen *Phyllachora maydis*. *P. maydis* is an ascomycete fungus classified as an obligate pathogen (Valle-Torres et al., 2020). Obligate pathogens require live plant material to complete their life cycle and to reproduce (Cline, 2005). The host range for *P. maydis* is specific to maize (Cline, 2005), but other species in the *Phyllachora* genus target a wide range of grass species (Parbery, 1967, 1971).

Tar spot of maize was first discovered in Mexico in 1904 and quickly spread throughout Latin and Central America (Bajet et al., 1994). It has been endemic to Latin America for many years (Bajet et al., 1994; Liu, 1973). In Mexico, yield losses of up to 58% have been documented on susceptible hybrids due to severe tar spot infection (Loladze et al., 2019). Tar spot was first discovered in the United States in 2015, and it has been rapidly spreading throughout the Midwest United States since, threatening the yield and plant health (Ruhl et al., 2016). The spread of the disease can be seen in figure 2.1, showing the first year of disease incidence in each county in the United States of America. Tar spot has currently been found in 19 total states and Ontario, Canada (Corn ipmPIPE, 2023). These states are mostly in the midwestern United States and the corn belt, but additional states are reporting tar spot incidence in the south. These states include Indiana, Illinois, Georgia, Pennsylvania, Minnesota, Florida, Iowa, Michigan, Wisconsin, Kentucky, Iowa, Ohio, New York, Missouri, Kansas, South Dakota, Virginia, and Maryland (;

Pandey et al., 2022; Collins et al., 2021; Malvick et al., 2020, Ruhl et al., 2016; McCoy et al., 2018; Onofre, 2022; Corn ipmPIPE, 2022).

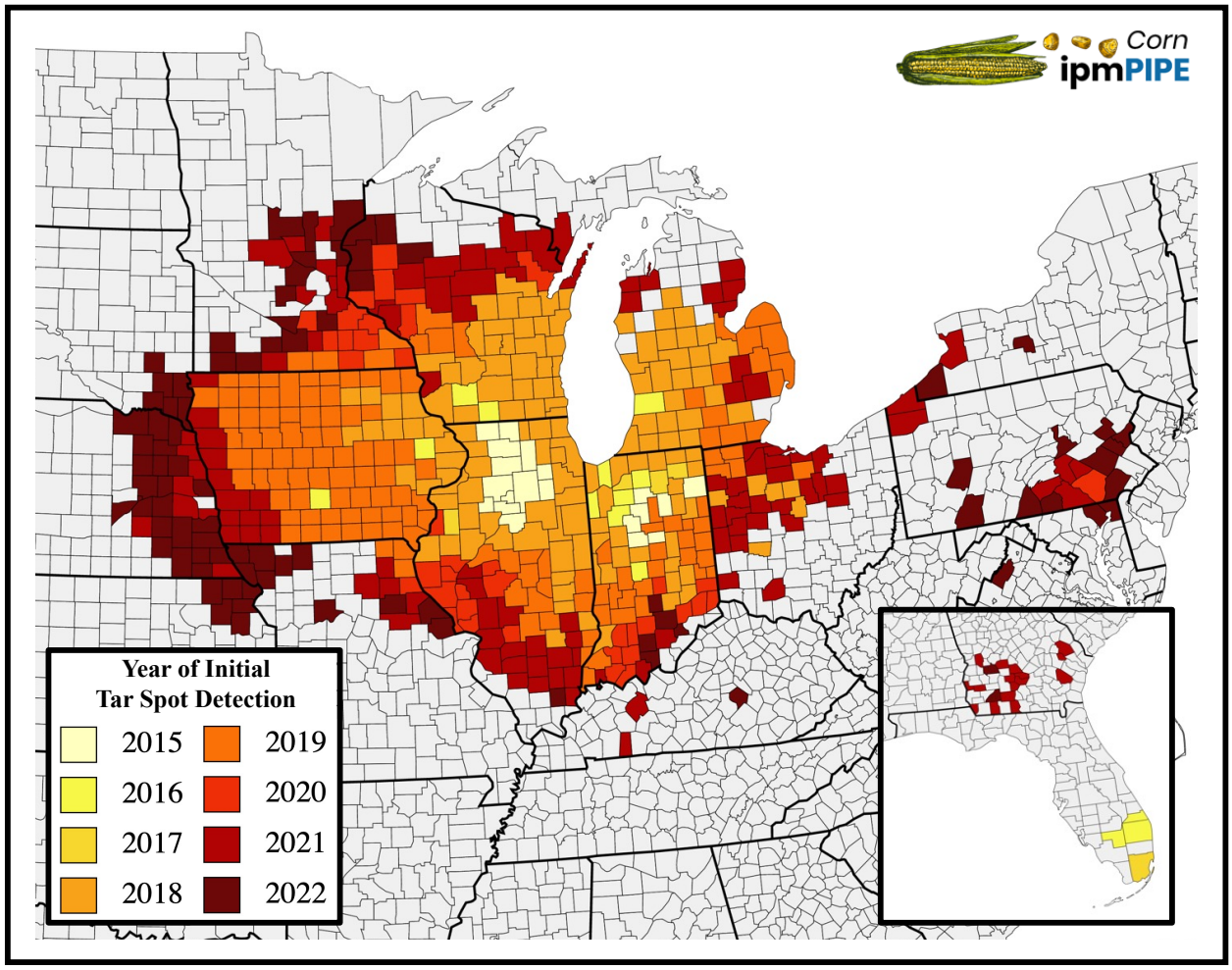


Figure 2.1: The year of initial detection of tar spot of maize in the United States of America. Color is used to indicate the years from 2015 to 2022. Disease incidence information sourced from Corn ipmPIPE (2022).

The primary inoculum for *P. maydis* is thought to be infected residue from the previous field season (Groves et al., 2020), or spores blown in from other nearby fields (Hock et al., 1992). The spores travel through water droplet splash and wind, creating the ability for the spores to spread to nearby plants and fields to cause infection (Hock et al., 1992; Valle-Torres et al., 2020). If the primary inoculum for *P. maydis* infection is from the overwintering spores on

plant residue, symptoms normally start with the lower leaves, then spread to the rest of the foliage (Cline, 2005). If the spores travel by wind or water splash, the symptoms can be found on upper leaves first (Cline, 2005). Tar spot of maize is identified by the formation of dense black stromata on the foliar tissues (Telenko & Creswell, 2019). The lesions are raised, shiny, and can permeate through the leaf; the stromata resemble tiny bits of tar and they cannot be removed from the leaf mechanically (Telenko & Creswell, 2019). In severe cases, the black stromata are encased in a brown halo lesion of necrotic tissue called fish-eye lesions (Cline, 2005). Over time, the halos, or fish-eye lesions, coalesce and the entirety of the leaf becomes chlorotic and necrotic (Yan et al., 2022). This can lead to rapid canopy senescence with severe infections (Telenko et al., 2019). The fisheye lesions are not common in the United States, but they can lead to sever yield losses in Central and Latin America. Originally, the fisheye lesions were thought to be caused by a separate pathogen, *Monographella maydis*, but research on the secondary pathogen is inconclusive and mostly done based on morphology and not sequencing (Mueller & Samuels, 1984). High disease severity and early season infection of *P. maydis* can cause up to 75% grain yield loss (Hock et al., 1989).

The complete understanding of the disease cycle of tar spot and its mechanisms are still ongoing due to its complex polycyclic nature. A polycyclic disease cycle produces multiple rounds of inoculum in a single growing season, allowing the reinfection of the plant numerous times (Bajet et al., 1994; Hock et al., 1989). The polycyclic nature of tar spot causes difficulties with disease management. Current management practices for tar spot are limited but include a range of chemical and cultural control measures. Cultural practices that are thought to help the spread and infection of tar spot disease include tilling, irrigation control, and timing planting (Telenko et al., 2020). Tilling fields with a known history of tar spot can promote the

decomposition of leftover plant tissue and minimize the survival of spores to the next growing season. This technique may help to reduce the primary inoculum for the next growing season, but it does not aid in the control of spores from nearby locations. Limiting excess irrigation practices of fields can also help reduce the severity of tar spot. This disease thrives in cool, wet, humid environments, and limiting excess irrigation treatments reduce the leaf wetness and prevent the optimal conditions for infection and pathogen spread (Hock et al., 1995; Valle-Torres et al., 2020). Chemical control measures have shown promise in preventing infection and reducing disease severity (Telenko et al., 2020). Fungicide applications can suppress tar spot, but it can be costly, need specialized equipment, and can cause potential harm to the surrounding environment. Sprayers effective on tall, mature maize are needed for proper application, but strategies change based on the specifics of the field history, plant varieties, and disease severity.

The current disease management practices for tar spot have a lot of room for improvement and are unable to completely prevent the disease. Even using a combination of chemical and cultural methods does not prevent the disease or provide farmers with the confidence that their yield will be protected from the devastating disease. In 2018, the United States experienced severe cases of tar spot infection causing almost 5 million metric tons in yield loss which equals over 680 million USD in economic losses (Mueller et al., 2020). Then, three years later, in 2021, the United States was faced with another year of severe tar spot infection causing grain yield losses of 5.88 million metric tons and an economic loss of 1.25 billion USD (Crop Protection Network, 2023).

Genetic resistance is the ability of the plant to be tolerant or resistant to the pathogen and the devastating effects of disease infection based on the genes alone, not including the addition of cultural or chemical control methods (Bent, 1996). Uncovering and exploiting genetic

resistance to tar spot of maize will prevent *P. maydis* from inoculating susceptible tissue and results in no or reduced infection of the plant. Genetic resistance of tar spot disease provides a cost-effective means of prevention without the harmful fungicides, expensive equipment, and difficult cultural practices.

Even though tar spot disease is relatively new to the United States, previous research has been done to uncover the genetic mechanisms behind potential disease resistance. A research study done by Mahuku et al. (2016) looked at 890 maize inbreds representing tropical and subtropical genetic diversity over multiple years in multiple environments. This study used low density markers for QTL mapping with an inclusive composite interval mapping technique to uncover a major QTL (*qRtsc8-1*) located at bin 8.03 connected to resistance to tar spot disease (Mahuku et al., 2016). Another research study, completed by Cao et al. (2017), aimed to dissect the genetic architecture of tar spot resistance in maize using an association mapping panel and three biparental doubled-haploid (DH) populations. This study found 4 different QTL, including one on chromosome 8 (bin 8.03), controlled tar spot disease resistance (Cao et al., 2017). Another study, performed by Yan et al. (2022), found 5 significant QTL regions through a mixed linear modeling genome wide association study technique. This study found that 172 significant SNPs were located in bin 8.03, along with additional significant SNPs found in bins 2.02, 3.03, 3.06, and 10.05 (Yan et al., 2022).

With the corresponding results found by Mahuku et al. (2016), Cao et al. (2017), and Yan et al. (2022), there was additional research interest on validating this major QTL through a fine mapping technique (Ren et al., 2022). The goal of the study completed by Ren et al. (2022), was to verify the effects of resistance markers which will speed up the breeding of elite varieties with genetic resistance to tar spot. Ren et al. (2022) fine mapped *qRtsc8-1* and identified the two

candidate genes *GRMZM2G063511* and *GRMZM2G073884*. The two candidate genes found have a high likelihood to be involved in the maize disease resistance response.

Another research study looked at over 600 varieties from the Wisconsin Diversity panel and 200 varieties from Iowa State's Germplasm Enhancement of Maize program to screen genetic resistance to tar spot (Trygestad, 2021). The research done by Trygestad (2021) uncovered over 100 significant SNPs connected with tar spot resistance and linked with candidate genes. None of the significant SNPs found by Trygestad (2021) were previously identified in tropical maize germplasm by Cao et al. (2017). The differences in results between Trygestad et al. (unpublished) and Mahuku et al. (2016), Cao et al. (2017), and Yan et al. (2022) allow us to understand that the complete mechanism to genetic resistance of tar spot is still not known. There is likely more minor effect QTLs to be found that will add to a higher level of resistance when used for breeding. New resistance QTL should be researched in different environments and populations to uncover additional useful genetic control elements in a wide variety of plant materials and environments.

Multi-parent Advanced Generation InterCrosses (MAGIC) breeding designs are considered useful in Dell'Acqua et al., 2015). Using a MAGIC population for an experiment creates genetic tools that are beneficial for mapping, have a high resolution, and create a high-power method for detecting QTL (Dell'Acqua et al., 2015). This structured MAGIC population will allow for a powerful QTL mapping structure due to the increased genetic diversity that the additional founding genomes provide (Scott et al., 2020). Exploring genetic resistance to tar spot in a MAGIC population could provide insight on new genes of interest and validate previously found results

This research study will utilize a MAGIC population with 6 founders (B73, B84, NKH8431, LH145, PHB47, and PHJ40) and 500 varieties to map the genetic resistance to tar spot of maize (Michel et al., 2022). The goals of this research study are to validate previously found resistance QTL, explore new QTL, exploit the increased power of using a MAGIC population, determine the parentage of resistance alleles, and discover any additional disease resistance QTL. The results found in this study will be used to increase understanding on the mechanisms behind genetic disease resistance for future breeding purposes to create elite varieties with genetic resistance to tar spot.

MATERIALS AND METHODS

Plant Material

Maize (*Zea mays*) germplasm is popularly classified into 3 main heterotic groups, stiff-stalk, non-stiff-stalk, and iodent. The stiff-stalk heterotic group was derived from the Iowa Stiff Stalk Synthetic (BSSS) created by Dr. George Sprague at Iowa State University. Commercial varieties of maize depend on the stiff-stalk germplasm for the use of high yielding hybrid varieties. Stiff stalk varieties were originally grouped based on high kernel yield and a smaller tassel, but the heterotic grouping of maize evolves as the varieties evolve.

The population used in this study is made up of 6 inbred lines that represent the range of diversity in the maize stiff stalk heterotic pool (Michel et al., 2022). The 6 stiff-stalk founder inbreds are B73, B84, NKH8431, LH145, PHB47, and PHJ40 (Michel et al., 2022). These founders were chosen to represent the range of diversity in the stiff-stalk heterotic group (Bornowski et al., 2022; Michel et al., 2022). The founders B73 and B84 originated at Iowa State University, and both are a part of the B73 sub-heterotic group (White et al., 2020). The founding

lines LH145 and NKH8431 originated from Holden's Foundation Seed, Inc. and Northrup, King & Company respectively, and are a part of the B14 sub-heterotic group (White et al., 2020). PHB47 and PHJ40 originated from Pioneer Hi-Bred International, Inc. and are a part of the B37 and Flint sub-heterotic group respectively (White et al., 2020). This population was created through multiparent advanced generation intercross (MAGIC) breeding design. 500 lines were derived from the population named WI-SS-MAGIC (Michel et al., 2022). The full MAGIC crossing scheme to derive the 500 lines used in this experiment can be found in Michel et al. (2022) along with more information about the population development process.

Experimental Design

The experiment was conducted in 3 different locations, Decatur, Michigan; Arlington, Wisconsin; and Washington Township, Indiana. The maize was planted in two-row plots with two replications planted in a randomized complete block design. The plots had a row spacing of 30 inches, a length of 10' with 3' alleys, and a planting density of 6 inches. The Michigan field was planted on May 11, 2021, the Wisconsin field was planted on May 18, 2021, and the Indiana field was planted on May 27, 2021. Experiment locations were selected based on fields having a known history of tar spot disease with high disease pressure (Telenko et al., 2020; Groves et al., 2020). Fields were irrigated to supplement rainfall and increase field moisture and to prolong leaf wetness, which is conducive to disease spread and considered an optimal environment for *P. maydis* (Telenko et al., 2020).

Phenotypic Evaluation

Each plot is assigned a single disease severity rating to represent the average disease severity of the plot. A numerical value is assigned between the values 0 and 100, 0 represents a plot with no disease infection or 0 % lesion coverage on the ear leaf, and 100 would represent 100% of the leaf covered by lesions. The severity rating is calculated from the average percent lesion coverage on 5 representative ear leaves. If there were any lesions in the entirety of the plot, even if it was not on an ear leaf, a value of 0.1 was given to show disease incidence. The disease severity rating scale images supplied by the Crop Protection Network are used as guides to assign percent lesion coverage and used as a method of standardization between those who collected the phenotypic data. The guide used for disease ratings can be seen in figure 2.2.

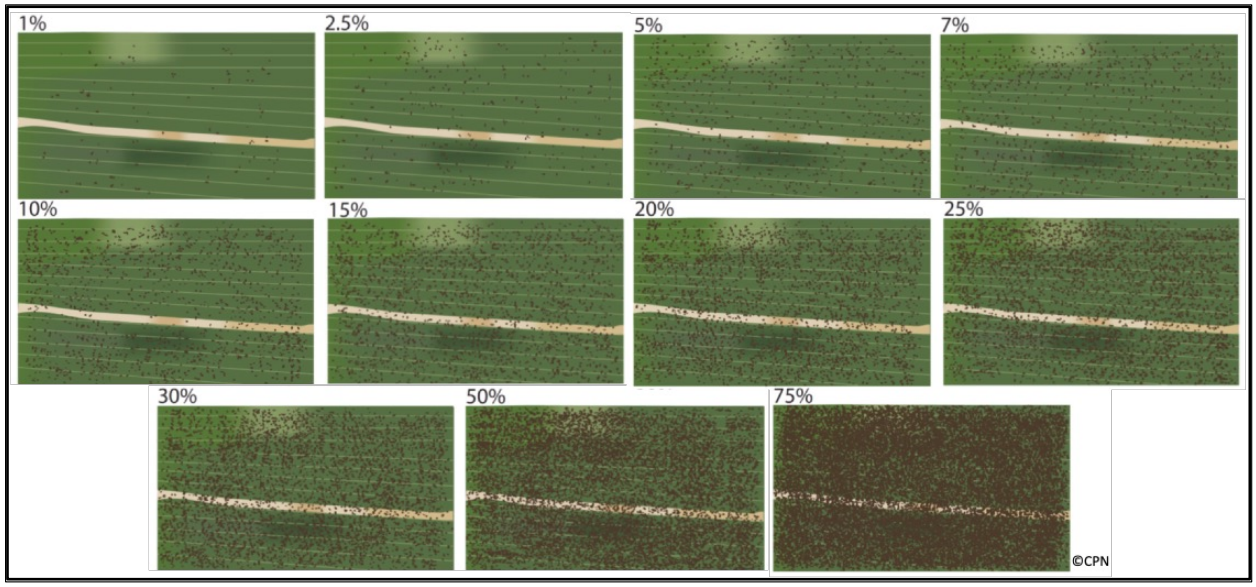


Figure 2.2: Tar spot disease severity rating scale based on percent severity. Percent severity is derived from the total stroma coverage of the black lesions on the ear leaf tissue. Images provided by the Crop Protection Network.

The disease severity ratings at the Michigan location began as soon as the first symptoms of tar spot presented. The ratings then continued as the disease progressed throughout the field

season, until leaf senescence began. The Michigan location performed ratings on July 19th, 2021, August 4th, 2021, August 18th, 2021, August 25th, 2021, and August 30th, 2021. Both the Wisconsin and Indiana locations were only rated once and it occurred at the end of the growing season, before the leaves began to dry down. The raw severity ratings for each field location can be found in the supplemental files.

Phenotypic Data Analysis

All statistical analysis and computational elements of this research are completed with R statistical software (R Core Team, 2022). Area under the disease progress curve (AUDPC) was calculated with the data collected in the Michigan location that had at least 4 of the 5 severity ratings. If there were additional missing ratings, it was deemed that the AUDPC would not be a quality representation of the disease progress in that plot. AUDPC was calculated using the R package agricolae (de Mendiburu & Yaseen, 2020). This R function uses the methodology of Campbell & Madden's 1990 publication, where t_i is the specific time in a sequence, y_i is the associated disease level at that time, and $y(0)$ is defined as the initial infection or the disease level at t = 0.

$$AUDPC = \sum_{i=1}^{n-1} \left(\frac{y_i + y_{i+1}}{2}\right)(t_{i+1} - t_i)$$

An analysis of variance (ANOVA) was performed on the tar spot severity data using the linear model function in R. The ANOVA was conducted to determine if there was significant difference between any of the genotypes and relative position within the field. Both genotype and relative field position were considered to be significant in all 3 of the locations with p-values all <0.0001. Field position in each of the locations was allocated by range and pass. An

additional ANOVA was done for each location using the linear model function in R using pass and range to predict the final tar spot severity ratings.

In order to account for the variance between the different researchers collecting disease ratings and the variance in the field, a moving grid adjustment was made using the mvngGrAd package in R (Technow, 2015). This package calculates the moving mean of the i^{th}entry, written as x_i and calculated as:

$$x_i = \frac{\Sigma_j \, p_{j,obs} \cdot I(p_{j,\ obs} \in G_i)}{\Sigma_j \, I(p_{j,\ obs} \in G_i)}$$

The value calculated for x_i is used as a covariate to calculate an adjusted phenotypic value ($p_{i,adj}$) through the use of the observed phenotype ($p_{i,obs}$), the moving mean x_i, and the overall mean with the formula:

$$p_{i,adj} = p_{i,obs} - b(x_i - \bar{x})$$

The adjusted phenotypic value takes into account 3 plots disease severity in every direction to help remove the variance based on placement. The selection of plots to include in the moving mean calculation can be adjusted for field and researcher preference.

QTL Analysis

The QTL analysis was completed using the R package qtl2 (Broman et al., 2019). The genotype data file included 100,000 SNPs coded with 0 and 1's using only genotyping by sequencing (GBS) methods. The number of meioses for the lines was captured using the crossing information, which was also used to create the population in qtl2. The physical and genetic map were created from the numeric genotype files and a conversion from Mbp to cM was done considering the total chromosomal length of 2.132Gbp and the total US-NAM length of 1,456.68

cM (https://www.maizegdb.org/). More detailed information on the sequencing methods and the creation of the genetic and physical map can be seen in Michel et al. (2022).

The marker data, with an assumed genotyping error probability of 0.01, was used to calculate the probabilities of the underlying genotypes. The genotype probabilities were then used to clean the data based on the number of crossovers and any duplicates. Individuals were removed if they had either 3 meioses and greater than 150 crossovers, or 5 meioses and greater than 250 crossovers. Additional information on the process of filtering the genotype data can be seen in Michel et al. (2022). Individuals that had marker sharing of greater than or equal to 0.95 were labeled as duplicates and removed. Next, a kinship matrix was calculated using 'loco' method, leaving one chromosome out at a time. To calculate the kinship matrix, the genotype probabilities were first converted to allele probabilities. Then, a full genome scan of all the phenotypes, including all three locations, raw values, adjusted values, and averages, were done using a restricted maximum likelihood method and least squares model. The QTL peaks were calculated using the find peaks function and a permutation test with 1000 permutations as the threshold; separate permutation tests were done for each phenotype. A probability of 0.95 and a peak drop of 5 was used as significance thresholds for mapping purposes.

Next, the BLUP effects were estimated on a chromosome basis for each of the phenotypes using the scan1blup function in r/qtl2 (Broman et al., 2019). This function calculates BLUPs of QTL effects along a chromosome, with a single QTL model while considering the QTL effects as random. The scan1blup function was used instead of the scan1coef function due to the multiple parental genotypes possible at a single QTL region. The BLUP effects allowed us to obtain the coefficient estimates by chromosome and sort them by each founder genotype. Once the coefficients were assigned to their respective founder genotype, the QTL BLUP effects

were visualized and traced back to specific six founding genotypes. The code used to perform the calculations, QTL analysis, and figure creation can be found in the supplemental files.

RESULTS

Disease Severity by Location

Disease severity ratings were taken on 5 separate occasions in the Michigan location. The first rating, taken on July 19th, 2021, ranges from 0.00% to 0.10% severity coverage of the ear leaf averaged over the plot, with a mean rating of 0.0032%. The second rating, taken on August 4th, 2021, ranges from 0.00% to 0.25% disease severity, with a mean of 0.0271%. The third rating, taken on August 18th, 2021, ranged from 0.00% to 1.00% with a mean severity of 0.1018%. The fourth rating, taken on August 25th, 2021, ranged from 0.00% to 7.00% with a mean rating of 0.3388%. The final rating was taken on August 30th, and ranged from 0.00% to 7.50% disease severity, with a mean rating of 0.7562%. At the Indiana location, disease ratings were taken once at the end of the growing season, but before the leaves began to senesce. The minimum ratings given was 0.00%, the maximum rating was 11.67%, and the mean rating was 3.608%. The Wisconsin location was also evaluated a single time at the end of the growing season, before the leaves began to senesce. The disease severity values ranged from 0.060% to 36.00% with a mean severity rating of 9.400%. A summary of the final disease severity ratings can be seen in table 2.1.

Table 2.1: Final tar spot severity ratings at each of the three locations. Severity is measured by the percent of black lesion coverage on the ear leaf, on average of each plot.

Michigan	
Min	0.0 %
Median	0.5 %
Max	7.5 %
Indiana	
Min	0.0 %
Median	3.0 %
Max	27.5%
Wisconsin	
Min	0.0 %
Median	5.7%
Max	36.0 %

The disease severity was greatest in Wisconsin, with a higher mean and maximum rating, while Michigan showed the smallest amount of disease symptoms. The three locations were all planted in fields with a history of tar spot disease, but with different environmental characteristics. Weather is a huge factor in disease severity and transmission, and it could explain the differences between the locations. A visual summary of the final disease ratings and their spread can be seen in figure 2.3.

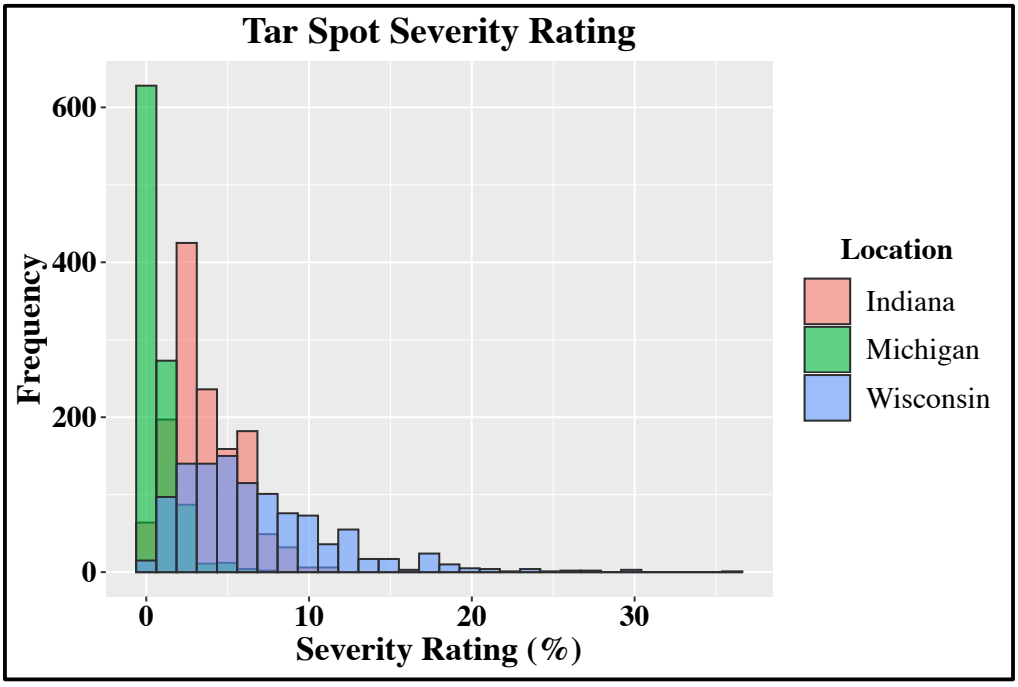


Figure 2.3: Frequency distribution of the final tar spot severity. Color indicates the location the ratings took place, Indiana is red, Michigan is green, and Wisconsin is blue.

Phenotypic Evaluation

Evaluating the disease severities for each line included an analysis of field position in comparison to the phenotypic disease severity value assigned. Tar spot is a disease that depends on the environment as a factor for the transmission and severity. Using a type III ANOVA to determine if the disease severity is affected by the range and pass location of the plot. For the Michigan location, both pass and range were considered significant with *p* values less than 0.0001. For the Indiana location, pass was significant with *p* values of <0.0001 and range was not significant with a *p* value of 0.0949. Lastly, for the Wisconsin location, pass was significant with a *p* value of less than 0.0001 and range was significant with a *p* value of 0.0002.

When calculating the environmental adjustments using the R package mvngGrAd, a total of three plots in every direction was used as a way to determine the moving mean (Technow, 2015). After the environmental adjustments were calculated, another ANOVA was calculated

using the adjusted phenotypic values for disease severity. The pass and range location were deemed not significant at the Michigan location with p-values of 0.3103 and 0.9872 respectively. A field graph, made with the R package desplot (Wright, 2021), showing the severity ratings before and after the environmental adjustments can be seen in figure 2.4. The Indiana location also resulted in pass and range values that were not significant with *p* values of 0.0562 and 0.9291 respectively. Lastly, after the environmental adjustment, Wisconsin's pass and range indicators for field position were not significant with p-values of 0.7861 and 0.9917 respectively. The phenotype file created with each of the locations data averaged per line can be found in the supplemental files.

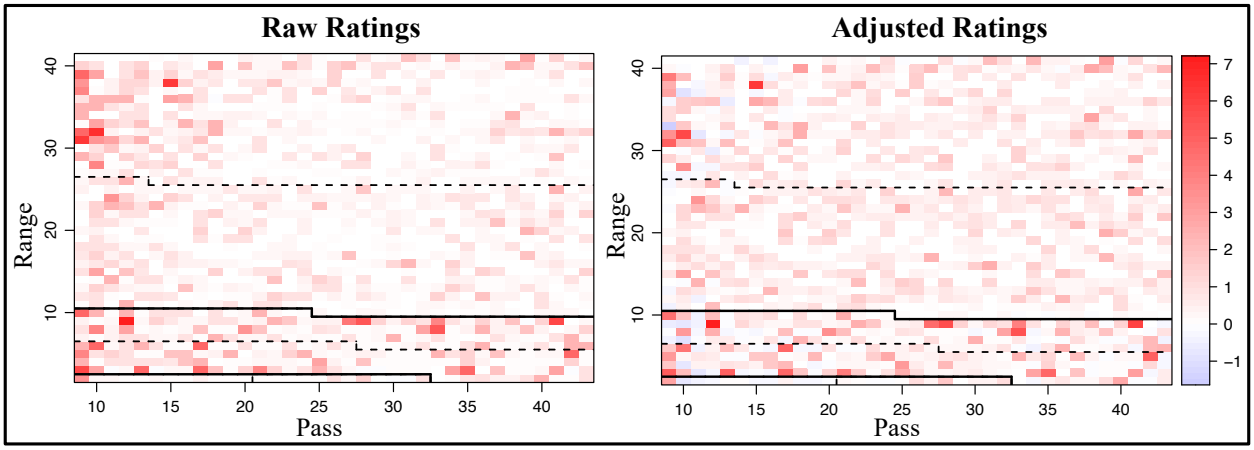


Figure 2.4: Field maps showing the tar spot severity ratings before and after the environmental adjustments. Red color indicates a positive severity, blue color indicates a negative severity, solid lines represent experimental blocks, and dashed lines represent replications within experimental blocks. Figures made through the R package desplot (Wright, 2021).

QTL Analysis

A permutation test was used to determine a threshold for a statistical test that determines if a result is due to random chance. The permutation test was completed with 1000 permutations to create a significance threshold with a *p* value of 0.05 for the LOD scores of the QTL analysis.

The significance thresholds calculated from the permutation test can be seen in table 2.2. The Michigan phenotypes received significance thresholds from the permutation test of 5.1398, 3.3571, 5.7091 for the raw values, environmentally adjusted values, and the AUDPC values respectively. The Indiana phenotypes received significance thresholds of 5.0311 for the raw values and 5.0098 for the environmentally adjusted values. The Wisconsin phenotypes received thresholds of 4.6552 and 4.4109 for the raw and environmentally adjusted values respectively. Lastly, the averaged raw values including the three locations received a significance threshold of 4.2300 and the averaged adjusted values received 3.8481 as a threshold.

Table 2.2: Significance thresholds for each phenotype determined by a permutation test with 1000 permutations. Permutation thresholds calculated using the scan1perm function from the R package qtl2 (Broman et al., 2019).

Trait	Permutation Threshold
MI_TS	5.1398
MI_TS_adj	3.3571
MI_TS_AUDPC	5.7091
IN_TS	5.0311
IN_TS_adj	5.0098
WI_TS	4.6552
WI_TS_adj	4.4109
TS_AVG	4.23
TS_AVG_Adj	3.8481

The raw Michigan disease severity ratings had significant QTL peaks on chromosome 1 and 9. The adjusted Michigan disease severity ratings had significant QTL peaks on chromosomes 1, 2, 3, 4, 5, 6, 7, 9, and 10. The derived AUDPC ratings from the Michigan disease severity ratings uncovered a single QTL peak on chromosome 1. The raw Indiana disease severity ratings showed significant QTL peaks on chromosome 1 and 8. The environmentally adjusted disease severity ratings from Indiana found one QTL peak on chromosome 1 and two QTL peaks on chromosome 8. The raw Wisconsin disease severity ratings showed significant

QTL peaks on chromosomes 1, 2, 4, and 8. The environmentally adjusted severity ratings from the Wisconsin location showed significant QTLs on chromosomes 1, 2, 4, and 6. The average of the raw severity ratings between the 3 locations showed significant QTLs on chromosomes 1, 2, 3, 4, and 8. A graph of the LOD scores plotted by position against the significance threshold can be seen in figure 2.5. The average of the environmentally adjusted disease severity ratings between the three locations showed QTL peaks on chromosomes 1, 2, 3, 4, 5, and 6. A summary of the QTL peaks and their LOD score can be seen in table 2.3.

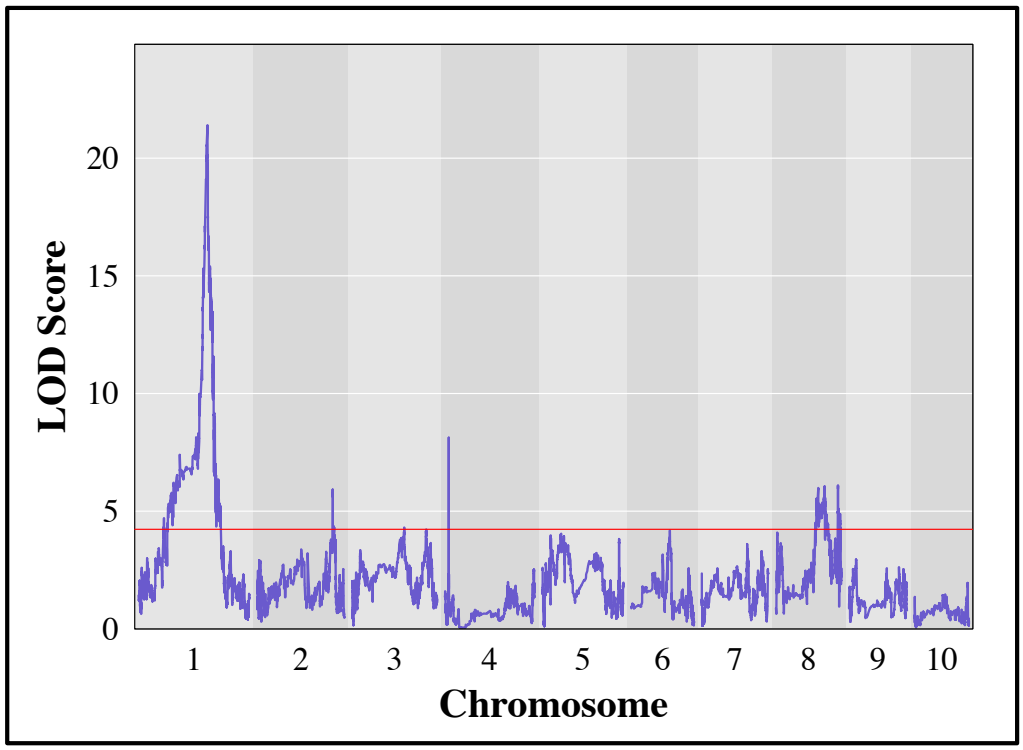


Figure 2.5: LOD scores by genome position for the averaged tar spot severity scores for all three experimental locations. The red line represents the significance threshold of 4.2300, calculated using a permutation test with 1000 permutations. Significant QTL are located on chromosomes 1,2,3,4, and 8.

Although, many of the QTL uncovered were unique to the location or trait. There were a few QTL that were common throughout multiple traits or locations. On chromosome 1, there was

a QTL in common throughout the Michigan final severity rating and the Michigan adjusted severity ratings. This QTL was found at 1.86 Mb and had the largest LOD score for each of the individual traits. Another QTL found in common amongst multiple traits is located on chromosome 1 at the position 1.91 Mb. This QTL was found in the traits MI_AUDPC, IN_TS, IN_adj, WI_TS, WI_TS_adj, TS_AVG, and TS_AVG_Adj. This QTL had an astoundingly large LOD score ranging from 6.75 (MI_AUDPC) to 24.35 (WI_TS_adj). It should be noted, the only two traits that did not have the significant peak at 1.91 Mb had a peak at 1.86 Mb. The range of this peak provided by the confidence interval encompasses the location of 1.91 Mb. This could mean that all of the peaks either located at 1.86 or 1.91 Mb could be the same or showing the significance of the same resistance element. Additional research is needed to determine the genetic architecture of that QTL.

Table 2.3: List of all of the QTL peaks found for each of the raw and calculated phenotypes at all three locations. The shortened trait names are MI_TS = Michigan final tar spot, MI_TS_adj = Michigan adjusted final tar spot, MI_AUDPC = Michigan area under disease progress curve, IN_TS = Indiana final rating, IN_TS_adj = Indiana final adjusted rating, WI_TS = Wisconsin final rating, WI_TS_adj = Wisconsin final adjusted rating, TS_AVG = averaged raw data for all locations, and TS_Adj_AVG = averaged adjusted values for all locations.

Trait	Chromosome	Position	LOD	CI_Low	CI_High
MI_TS	1	186959204	6.083	184051304	191440220
MI_TS	9	101500956	5.142	29380397	112919352
MI_TS_adj	1	186959204	7.104	184051304	192290238
MI_TS_adj	2	9348038	3.76	837221	243129967
MI_TS_adj	3	162681201	4.253	1213108	215318144
MI_TS_adj	4	32364990	4.774	18776013	39070571
MI_TS_adj	5	13796401	3.908	287665	214951844
MI_TS_adj	6	117670369	3.393	47955050	180388827
MI_TS_adj	7	109546473	3.594	15591043	185225465
MI_TS_adj	9	103067265	3.583	2056815	162741970
MI_TS_adj	10	146614609	3.488	1253208	151594788
MI_AUDPC	1	191290617	6.751	184010012	208287420
IN_TS	1	191290617	16.519	188403119	191866073
IN_TS	8	112707107	6.638	105283076	173240428
IN_TS_adj	1	191291342	16.525	188098266	191631139
IN_TS_adj	8	112707107	6.534	105283076	142217080
IN_TS_adj	8	172347649	5.589	163971439	177304637
WI_TS	1	191289217	23.997	188405829	191440220
WI_TS	2	210212746	6.694	205523753	215041419
WI_TS	4	11862714	5.22	1034184	250055869
WI_TS	8	172349578	5.693	3971943	178987840
WI_TS_adj	1	191289332	24.348	188405829	191440220
WI_TS_adj	2	210212606	5.362	67741681	218191247
WI_TS_adj	4	11864631	4.704	1034184	250055869
WI_TS_adj	6	112859912	4.759	105319502	117624413
WI_TS_adj	8	172499685	6.891	115748200	178962163
TS_AVG	1	191289332	21.402	188405829	191440220
TS_AVG	2	210212746	5.918	122494603	218293623
TS_AVG	3	145540246	4.307	12115644	223093363
TS_AVG	4	11871181	8.124	11863028	11871460
TS_AVG	8	172893982	6.091	112460863	178483320
TS_Adj_AVG	1	191289332	21.082	188403119	191440220
TS_Adj_AVG	2	210212746	4.74	64584851	239612041
TS_Adj_AVG	3	145540246	4.898	12115644	222598227
TS_Adj_AVG	4	11871181	7.661	1778603	249333396
TS_Adj_AVG	5	23536351	4.195	287665	225487769
TS_Adj_AVG	6	112860189	4.656	64898348	167090109
TS_Adj_AVG	8	172368251	6.971	112066603	178434028

QTL Effects

Using R/qtl2 (Broman et al., 2021), the BLUP effects were estimated for each of the founders of the MAGIC population per chromosome and separately for each of the phenotypes. The complete data tables for each phenotype and each chromosome can be found in the supplemental files. This method helps dissect the variability at each QTL due to the numerous possible genotypes in this population. Inspecting the large QTL on chromosome 1 displays an obvious divide in the founders that are causing a positive or negative QTL effect. The plotted results can be seen in figure 2.6 where each of the lines on the plot is the QTL effects due to the specific founder. Four of the founder genotypes, NKH8431, LH145, PHB47, and PHJ40, show a positive QTL effect, which indicates lower resistance, or higher disease severity.

The 2 founder genotypes, B73 and B84, are correlated to a negative QTL effect. A negative QTL effect implies a greater amount of resistance or lower disease severity. Varieties with B73 and B84 parentage at specific locations on chromosome 1 display about 1.25% lower disease severity than the other founders. Coupling alleles from those two founders at multiple points on chromosome 1 may display a larger amount of resistance. Using the QTL effects split by founders for each chromosome can show a combination of alleles from different parentages that could lead to lower severity and heightened resistance for future breeding efforts.

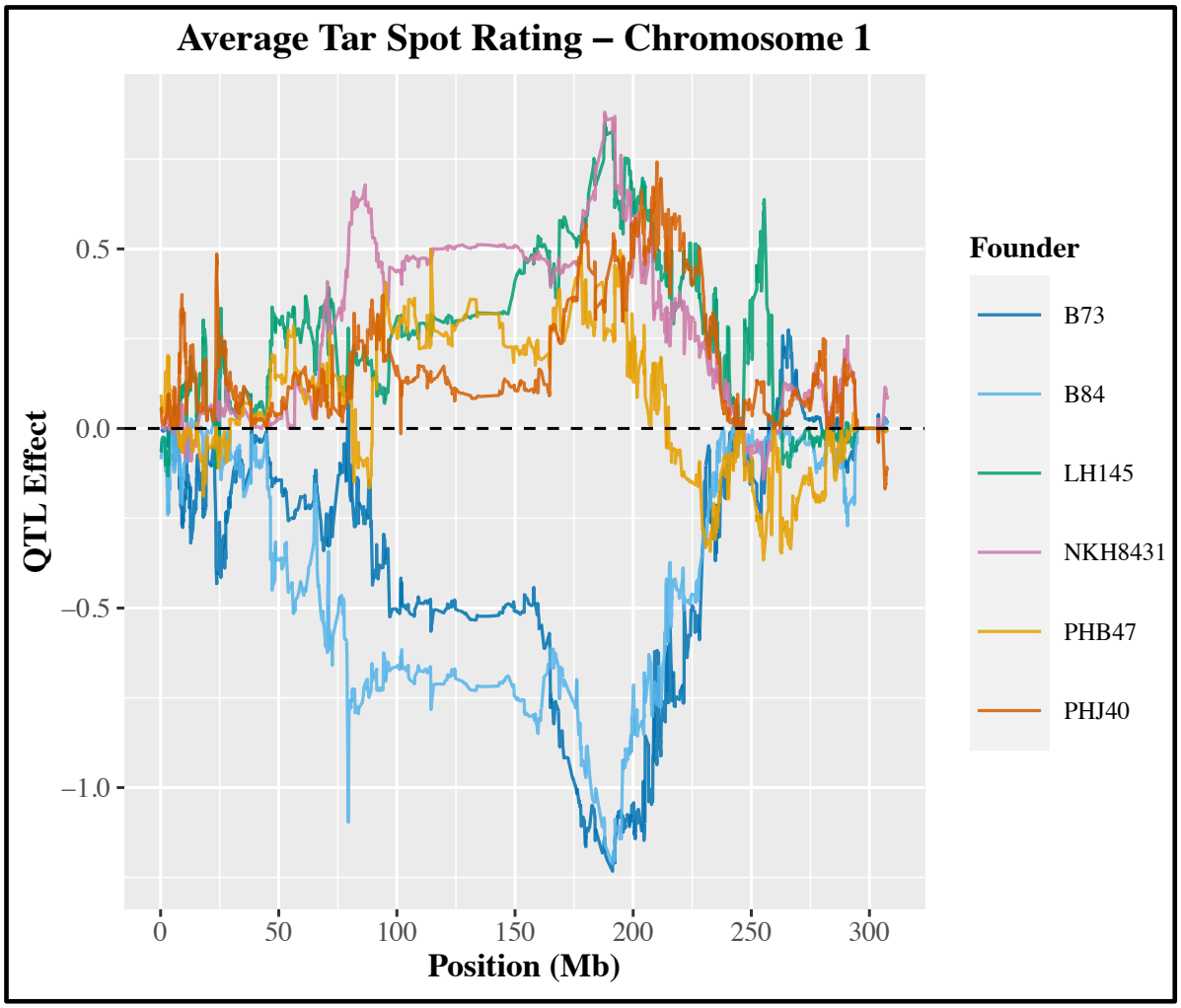


Figure 2.6: QTL effects on the average tar spot severity ratings on chromosome one, separated by population founder. Color indicates the six different population founders. A negative QTL effect correlates to lower disease severity and higher resistance. A positive QTL effect correlates to a higher disease severity and lower resistance.

DISCUSSION

Environmental effects on tar spot

Tar spot severity and spread is highly dependent on the environmental conditions. The mode of initial infection depends on the soil and wind conditions. Means of secondary inoculum infection depends on wind and water splash, and severity is dependent on the optimal environmental conditions that are conducive to *P. maydis*. Obligate pathogens, like *P. maydis*, are incredible difficult to produce a lab inoculation technique because live plant material is

needed to complete the lifecycle. With no current means of artificial inoculation at the time this research is conducted, inoculation and severity are completely dependent on the presence of spores overwintered in the field, surrounding fields history of disease, and the environmental conditions.

With the Michigan location in particular, the first symptoms of disease were not present until mid-July due to the dry conditions in the spring and early summer. By the time disease severity started to increase, the plants began to dry down and senesce in preparation for harvest. The dry conditions caused a slow start to the disease spread which reduced the initial severity. The MAGIC population used for this study is also known to be early maturing. A combination of the early planting date, dry conditions, early maturing varieties, and late presence of disease symptoms, the severity was lower than the other two locations. With the Michigan location, we may have seen different results if we were able to take additional ratings before the drying down of the leaves, or if the symptoms of infection were present sooner. This is one of the limitations of field experiments on a disease without artificial inoculation techniques. The Indiana and Wisconsin field did not see these environmental issues to the same degree that the Michigan field did. The Indiana and Wisconsin fields saw higher disease pressure, greater severity, and were able to take their severity ratings in mid to late September before any leaf senescence occurred.

The development of an artificial inoculation technique is currently being studied by multiple researchers (Breunig et al., 2023; Góngora-Canul et al., 2023; Kleczewski et al., 2019). Multiple of these researchers have accomplished an artificial inoculation technique, but the disease severity has been low. The success of creating an inoculation technique for this obligate pathogen that would cause high disease severity could allow for genetic screening techniques

that are not limited by the environment. It could create a more standardized system and remove some of the variance caused by the current modes of disease infection and spread.

Implications of uncovered genetic resistance QTL

Using a stiff-stalk MAGIC population to map the genetic resistance to tar spot of maize has uncovered additional significant QTL. With previous research done by Mahuku et al. (2016), Cao et al. (2017), and Yan et al. (2022) focusing on the significant major QTL on chromosome 8 (bin 8.03), there is room for uncovering additional resistance QTL. Most quantitative traits are controlled by numerous small effects QTL that can be difficult to detect. The added power a MAGIC population can provide led to additional insight on more of these smaller effect resistance loci.

This research study was successful in uncovering additional significant resistance QTL with varied magnitude of effects. Although, the positions of these QTL will aid in future research and plant breeding efforts, additional investigation is needed to examine the QTL, the effects, and the surrounding genes. For instance, the QTL found on chromosome 1 located at about 1.91 Mb has an astonishingly large LOD when compared with the other significant QTL. This QTL also spans a large range and takes up a large area on the genome. With the research done in this study, we were unable to determine if this is a single large peak, or if there are multiple significant QTL peaks within the large range it spans. With a higher density map, or a different fine mapping technique used on chromosome 1, it may be possible to refine the peak range or number of significant peaks. This could reduce the number of potential candidate genes and highlight a smaller length of the genome that researchers can study.

Candidate genes within QTL peaks for resistance

Candidate genes are found within range of the significant QTL peak. Candidate genes are potential genes that could be causing the genetic resistance to tar spot in that location on the genome. The QTL peak has a range calculated around the significant point from the confidence interval where the resistance element may be in the genome. Using the B73v4 reference genome (https://www.maizegdb.org/) and the BEDTools package on command line, the genes that fall in that range were extracted and compiled (Quinlan & Hall, 2010). To remove some redundancies, QTL peaks that were within 25,000 bp of each other were merged and considered as a single peak. All of the traits are listed for the merged peak, and the range was taken from the original peak with the highest LOD score. With the above regulations, over 40,000 potential candidate genes were extracted. The full list of potential candidate genes can be found in supplemental files. Looking at the closest gene to each of the QTL peaks, we can reduce this number to 20 potential candidate genes. The list of the closest candidate genes to the QTL peak was found using MaizeGDB (Woodhouse et al., 2021). A summary of the significant QTL peaks, the phenotypes, and the candidate genes closest to the marker at the peak can be found in table 2.4.

Table 2.4: Candidate genes closest to the QTL peaks. QTL peaks among the list of phenotypes that are within 25,000 bp of each other are considered a single peak and all phenotypes where this peak is found are listed in the traits column. Genes were found using the B73 reference genome version 4.

Chr	Marker	Peak (bp)	Traits	Gene
1	Chr1_186959204	186959204	MI_TS MI_TS_adj	Zm00001d031279
1	Chr1_191289217	191289217	WI_TS WI_TS_adj TS_AVG TS_Adj_AVG MI_AUDPC IN_TS IN_TS_adj	Zm00001d031412
2	Chr2_9348038	9348038	MI_TS_adj	Zm00001d002259
2	Chr2_210212746	210212746	WI_TS TS_AVG WI_TS_adj TS_Adj_AVG	Zm00001d006507
3	Chr3_145540246	145540246	TS_AVG TS_Adj_AVG	Zm00001d041941
3	Chr3_162681201	162681201	MI_TS_adj	Zm00001d042340
4	Chr4_11871181	11871181	TS_AVG TS_Adj_AVG WI_TS WI_TS_adj	Zm00001d048990
4	Chr4_32364990	32364990	MI_TS_adj	Zm00001d049485
5	Chr5_13796401	13796401	MI_TS_adj	Zm00001d013527
5	Chr5_23536351	23536351	TS_Adj_AVG	Zm00001d013859
6	Chr6_112859912	112859912	WI_TS_adj TS_Adj_AVG	Zm00001d036896
6	Chr6_117670369	117670369	MI_TS_adj	Zm00001d037055
7	Chr7_109546473	109546473	MI_TS_adj	Zm00001d020345
8	Chr8_112707107	112707107	IN_TS IN_TS_adj	Zm00001d010398
8	Chr8_172368251	172368251	TS_Adj_AVG WI_TS IN_TS_adj	Zm00001d012255
8	Chr8_172499685	172499685	WI_TS_adj	Zm00001d012257
8	Chr8_172893982	172893982	TS_AVG	Zm00001d012267
9	Chr9_101500956	101500956	MI_TS	Zm00001d046627
9	Chr9_103067265	103067265	MI_TS_adj	Zm00001d046664
10	Chr10_146614609	146614609	MI_TS_adj	Zm00001d026397

Breeding for resistance to tar spot

Elite varieties of maize are continuously being bred for their stressor resistance, increased yield, and high farmer and consumer quality traits. There are currently maize hybrids that show some resistance to tar spot, but there is not a specific variety that is bred for complete tar spot resistance. The research done in this study highlighted numerous QTL found connected to the genetic mechanisms that control tar spot resistance. Determining the function of the candidate genes in the ranges of the QTL peaks through mutant studies will allow a better understanding of how the candidate genes attribute to the phenotypic variance. With a better understanding of the candidate genes and determining if they are strongly linked to a marker, plant breeders will be able to introgress these genes into their breeding programs to create new elite varieties with tar spot resistance. To do this, additional research needs to be done refining the QTL peak areas and narrowing down the number of potential candidate genes.

The results of the QTL effects for each founder show that the founders B73 and B84 have alleles that correlate to tar spot resistance. Introducing these two varieties into breeding programs will introduce the resistant alleles, especially found in the large peak on chromosome one, which will help decrease the severity of tar spot infection and promote overall resistance. Using the results in this study to add to the overall understanding of the genetic mechanisms that control tar spot resistance allows researchers and plant breeders to be one step closer to creating new elite resistant varieties of maize.

CONCLUSION

Currently, combatting tar spot of maize includes altering management practices and utilizing appropriate fungicides. While fungicides are able to reduce tar spot disease pressure and

counteract some of the devastating yield effects, management techniques should not be solely relied on chemical control. This research study used a MAGIC population to increase the power of a QTL analysis and uncover new resistance QTL and validate previously found genetic control mechanisms. Utilizing the discovered QTL and founder effects in the population to influence breeding decisions, will lead to the incorporation of resistance genes and alleles into breeding programs and new elite maize varieties. Implementing greater levels of genetic resistance will reduce the need for costly and potentially harmful chemical fungicide applications and will provide a stronger means of management for this complex disease.

While the research done in this study is helpful and can be utilized in future breeding programs, additional research will need to be done in order to fully understand the genetic mechanisms of tar spot resistance and to introgress these resistance QTL into elite varieties. The next steps of this study will look at two populations of doubled haploids made from crossing resistant lines to either B73 or LH244. Disease severity screenings will be done to explore the phenotypic variance and to further understand the passing of resistant alleles to offspring. The doubled haploid population with B73 as a parent will exploit previously found resistant lines in addition to the results of the founder effects from the stiff-stalk MAGIC population. More general next steps for furthering this research will be to refine the significant QTL peaks, narrow down the candidate genes, and examine the phenotypic variance of these genes in relation to tar spot infection.

Tar spot of maize is a fungal disease that has just begun causing devastating effects to our corn yield resulting in severe economic losses. It is predicted that tar spot infection will continue to spread across the US and Canada, continuing to increase. With a greater area of potential disease infection, a field season with optimal environmental conditions for *P. maydis* infection

will lead to substantial grain yield and economic losses. Using genetic resistance as a management technique for tar spot will help counteract the potential for the devastating effects.

REFERENCES

Bajet, N.B., Renfro, B.L., & Carrasco, J.M.V. (1994). Control of tar spot of maize and its effect on yield. *Int. J. Pest Manage*. 40:121-125. https://doi.org/10.1080/09670879409371868

Bent, A. F. (1996). Plant disease resistance genes: Function meets structure. *The Plant Cell*, 1757–1771. https://doi.org/10.1105/tpc.8.10.1757

Bornowski, N., Michel, K.J., Hamilton, J.P., Ou, S., Seetharam, A.S., Jenkins, J., Grimwood, J., Plott, C., Shu, S., Talag, J., Kennedy, M., Hundley, H., Singan, V.R., Barry, K., Daum, C., Yoshinaga, Y., Schmutz, J., Hirsch, C.N., Hufford, M.B., de Leon, N., Kaeppler S.M., & Buell, C.R. 2021. Genomic variation within the maize stiff-stalk heterotic germplasm pool. *The Plant Genome, 14*(3). **h**ttps://doi.org/10.1002/tpg2.20114

Breunig, M., Bittner, R., Dolezal, A., Ramcharan, A., & Bunkers, G. (2023). *An Assay to Reliably Achieve Tar Spot Symptoms on Corn in a Controlled Environment*. bioRxiv. https://doi.org/10.1101/2023.01.12.523803

Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Sen Ś, Yandell BS, Churchill GA (2019) R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multi-parent populations. Genetics 211:495-502

Campbell, C. L., L. V. Madden. (1990): Introduction to Plant Disease Epidemiology. John Wiley & Sons, New York City.

Cao, S., Loladze, A., Yuan, Y., Wu, Y., Zhang, A., Chen, J., Huestis, G., Cao, J., Chaikam, V., Olsen, M., Prasanna, B.M., San Vicente, F., & Zhang, X. 2017. Genome wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction. *The Plant Genome, 10*(2). https://doi.org/10.3835/plantgenome2016.10.0099

Cline, E. (2005). *Phyllachora maydis*. U.S. National Fungus Collections, ARS, USDA. Retrieved June 30, 2021, from https://nt.ars-grin.gov/sbmlweb/fungi/nomenSheets.cfm

Collins, A. A., Bandara, A. Y., May, S. R., Weerasooriya, D. K., & Esker, P. D. (2021). First report of tar spot of maize (*zea mays*) caused by *Phyllachora Maydis* in Pennsylvania. *Plant Disease*, *105*(8), 2244. https://doi.org/10.1094/pdis-11-20-2456-pdn

Corn ipmPIPE. (2022, April 29). *Tar spot*. Corn ipmPIPE. https://corn.ipmpipe.org/tarspot/

Corn ipmPIPE. (2023, February 23). Historical end of season maps. Tar Spot. https://corn.ipmpipe.org/tarspot/historical-end-of-season-maps/

Crop Protection Network (2023). Estimates of crop yield losses due to diseases and invertebrate pests: an online tool https://loss.cropprotectionnetwork.org/ https://doi.org/10.31274/cpn-20191121-0 (Accessed July 18, 2023).

Dell'Acqua, M., Gatti, D. M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., Hlaing, A. L., Aung, H. H., Neilissen, H., Baute, J., Frascaroli, E., Churchill, G., Inzé, D., Morgante, Michele, & Pè, M. E. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biology, 16*(167). https://doi.org/10.1186/s13059-015-0716-z

Góngora-Canul, C., Jiménez-Beitia, F. E., Puerto-Hernández, C., Avellaneda C., M. C., Kleczewski, N., Telenko, D. E., Shim, S., Solórzano, J. E., Goodwin, S. B., Scofield, S. R., & Cruz, C. D. (2023). Assessment of symptom induction via artificial inoculation of the obligate biotrophic fungus *Phyllachora maydis* (Maubl.) on corn leaves. *BMC Research Notes*, *16*(1). https://doi.org/10.1186/s13104-023-06341-y

Groves, G. L., Kleczewski, N. M., Telenko, D. E. P., Chilvers, M. I., & Smith, D. L. (2020). *Phyllachora maydis* Ascospore release and germination from overwintered corn residue. *Plant Health Progress*, *21*,1. https://doi.org/10.1094/PHP-10-19-0077-RS

Hock, J., Kranz, J., & Renfro, B. (1989). El complejo 'mancha de asfalto' de maíz: Su distribucción geográfica, requisitos ambientales e importancia económica en México. *Revista Mexicana de Fitopatologia*, 7(2), 129–35.

Hock, J., Dittrich, U., Renfro, B. L., & Kranz, J. (1992). Sequential development of pathogens in the maize tarspot disease complex. *Mycopathologia*, 117(3): 157–161. https://doi.org/10.1007/BF00442777

Hock, J., Kranz, J., & Renfro B. L. (1995). Studies on the epidemiology of the tar spot disease complex of maize in Mexico. *Plant Pathology*, 44(3):490–502. https://doi.org/10.1111/j.1365-3059.1995.tb01671.x

Liu, L. J. (1973). Incidence of tar spot disease of corn in Puerto Rico. *J. Agric. Univ. Puerto Rico*, *42*, 211-216.

Loladze, A., Rodrigues, F.A., Toledo, F., San Vicente, F., Gérard, B., & Boddupalli, M.P. (2019). Application of remote sensing for phenotyping tar spot complex resistance in maize. *Frontiers in Plant Science*, 10, 552. https://doi.org/10.3389/fpls.2019.00552

Kleczewski, N. M., Donnelly, J., & Higgins, R. (2019). *Phyllachora Maydis*, causal agent of tar spot on corn, can overwinter in Northern Illinois. *Plant Health Progress*, *20*(3), 178–178. https://doi.org/10.1094/php-04-19-0030-br

Mahuku, G., Chen, J., Shrestha, R., Narro, L., Guerrero, K. V. O., Arcos, A. L., and Xu, Y. (2016). Combined linkage and association mapping identifies a major QTL (*qRtsc8-1*), conferring tar spot complex resistance in maize. *Theoretical and Applied Genetics, 129*, 1217-1229. https://doi.org/10.1007/s00122-016-2698-y

Malvick, D. K., Plewa, D. E., Lara, D., Kleczewski, N. M., Floyd, C. M., & Arenz, B. E. (2020). First report of tar spot of corn caused by *Phyllachora Maydis* in Minnesota. *Plant Disease*, *104*(6), 1865–1865. https://doi.org/10.1094/pdis-10-19-2167-pdn

McCoy, A. G., Romberg, M. K., Zaworski, E. R., Robertson, A. E., Phibbs, A., Hudelson, B. D., Smith, D. L., Beiriger, R. L., Raid, R. N., Byrne, J. M., & Chilvers, M. I. (2018). First report of tar spot on corn (*zea mays*) caused by *Phyllachora Maydis* in Florida, Iowa, Michigan, and Wisconsin. *Plant Disease*, *102*(9), 1851. https://doi.org/10.1094/pdis-02-18-0271-pdn

de Mendiburu, F., & Yaseen, M. 2020. agricolae: Statistical Procudures for Agricultural Research.R package version1.4.0, https://myaseen208.github.io/agricolae/https://cran.r-project.org/package=agricolae.

Michel, K.J., Lima, D.C., Hundley, H., Singan, V., Yoshinaga, Y., Daum, C., Barry, K., Broman, K.W., Buell, C.R., de Leon, N., & Kaeppler, S.M. (2022). Genetic mapping and prediction of flowering time and plant height in a maize Stiff Stalk MAGIC population. *Genetics, 221*(2). https://doi.org/10.1093/genetics/iyac063

Mueller, D. S., Wise, K. A., Sisson, A. J., Allen, T. W., Bergstrom, G. C., Bissonnette, K. M., Bradley, C. A., Byamukama, E., Chilvers, M. I., Collins, A. A., Esker, P. D., Faske, T. R., Friskop, A. J., Hagan, A. K., Heiniger, R. W., Hollier, C. A., Isakeit, T., Jackson-Ziems, T. A., Jardine, D. J., & Wiebold, W. J. (2020). Corn yield loss estimates due to diseases in the United States and Ontario, Canada, from 2016 to 2019. *Plant Health Progress*, *21*(4), 238–247. https://doi.org/10.1094/php-05-20-0038-rs

Onofre, R. (2022, October 13). *Tar spot of corn is now confirmed in five counties in Kansas*. Agronomy eUpdate October 13th, 2022: Issue 928. https://eupdate.agronomy.ksu.edu/article_new/tar-spot-of-corn-is-now-confirmed-in-five-counties-in-kansas-516-5

Pandey, L., Burks, C. A., Gómez Londoño, L., Newsom, L., Brock, J. H., Kemerait, R. C., & Brewer, M. T. (2022). First report of tar spot on corn caused by *Phyllachora Maydis* in Georgia, United States. *Plant Disease*, *106*(8), 2262. https://doi.org/10.1094/pdis-11-21-2456-pdn

Parbery, D.G. (1967). Studies on graminicolous species of Phyllachora Nke. in Fckl. V. A taxonomic monograph. *Australian Journal of Botany, 15*, 271-375.

Parbery, D.G. (1971). Studies on Graminicolous species, of Phyllachora Nke. in Fckl. VI. Additions and corrections to part V. *Australian Journal of Botany, 19*, 207-235.

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

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ren, J., Wu, P., Huestis, G.M., Zhang, A., Qu, J., Liu, Y., Zheng, H., Alakonya, A.E., Dhliwayo, T., Olsen, M., San Vicente, F., Prasanna, B.M., Chen, J., & Zhang, X. 2022. Identification and fine mapping of a major QTL (*qRtsc8-1*) conferring resistance to maize tar spot complex and validation of production markers in breeding lines. *Theoretical and Applied Genetics, 135*, 1551-1563. https://doi.org/10.1007/s00122-022-04053-8

Rudolph, C. (2023, March 22). *MSU researchers identifying corn tar spot management strategies*. AgBioResearch. https://www.canr.msu.edu/news/msu-researchers-identifying-corn-tar-spot-management-strategies#:~:text=Prior%20to%202015%2C%20tar%20spot,begin%20to%20destroy%20plant%20tissue.

Ruhl, G., Romberg, M. K., Bissonnette, S., Plewa, D., Creswell, T., & Wise, K. A. (2016). First report of tar spot on corn caused by *Phyllachora Maydis* in the United States. *Plant Disease*, *100*(7), 1496. https://doi.org/10.1094/pdis-12-15-1506-pdn

Scott, M.F., Ladejobi, O., Amer, S., Bently A.R., Biernaskie, J., Boden, S.A., Clark, M., Dell'Acqua, M., Dixon, L.E., Filippi, C.V., Fradgley, N., Gardner, K.A., Mackay, I.J., O'Sullivan, D., Percival-Alwyn, L., Roorkiwas, M., Singh, R.K., Thudi, M., Varshney, R.K., Venturini, L., Whan, A., Cockram, J., & Mott, R. 2020. Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity* **125**, 396–416. https://doi.org/10.1038/s41437-020-0336-6

Tar spot. Corn ipmPIPE. (2022, April 29). https://corn.ipmpipe.org/tarspot/

Technow, F. 2015. R package mvngGrAd: moving grid adjustment in plant breeding field trials. R package version 0.1.5.

Telenko, D. E., Chilvers, M. I., Byrne, A. M., Check, J. C., Da Silva, C. R., Kleczewski, N. M., Roggenkamp, E. E., Ross, T. J., & Smith, D. L. (2022). Fungicide efficacy on tar spot and yield of corn in the Midwestern United States. *Plant Health Progress*, *23*(3), 281–287. https://doi.org/10.1094/php-10-21-0125-rs

Telenko, D., Chilvers, M. I., Kleczewski, N., Mueller, D., Plewa, D., Robertson, A., Smith, D., Tenuta, A., & Wise, K. (2020). An overview of tar spot. *Crop Protection Network*. https://doi.org/10.31274/cpn-20190620-008

Telenko, D. E. P., Chilvers, M. I., Kleczewski, N., Smith, D. L., Byrne, A. M., Devillez, P., Diallo, T., Higgins, R., Joos, D., Kohn, K., Lauer, J., Mueller, B., Singh, M. P., Widdicombe, W. D., & Williams, L. A. (2019). How tar spot of corn impacted hybrid yields during the 2018 midwest epidemic. *Crop Protection Network*. https://doi.org/10.31274/cpn-20190729-002

Telenko, D., & Creswell, T. (2019). *Diseases of corn: Tar spot*. Purdue Extension. https://www.extension.purdue.edu/extmedia/BP/BP-90-W.pdf

Trygestad, B. (2021). Genetic And Genetic by Environment Effects on Tar Spot Resistance and Hybrid Yield in Maize (thesis).

White, M.R., Mikel, M.A., Leon, N., Kaeppler, S.M. 2020. Diversity and heterotic patterns in North American proprietary dent maize germplasm. *Crop Science, 60*(1):100–114. doi:10.1002/csc2.20050.

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., Harper, L. C., Gardiner, J. M., Schaeffer, M. L., & Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol, 21*, 385. https://doi.org/10.1186/s12870-021-03173-5

Wright, K. (2021). desplot: Plotting Field Plans for Agricultural Experiments_. R package version 1.9, https://CRAN.R-project.org/package=desplot

Yan, S., Loladze, A., Wang, N., Sun, S., Chilvers, M. I., Olsen, M., Burgueño, J., Petroli, C. D., Molnar, T., San Vicente, F., Zhang, X., & Prasanna Boddupalli, M. (2022). Association mapping of resistance to tar spot complex in maize. *Plant Breeding*, *141*(6), 745–755. https://doi.org/10.1111/pbr.13056

CHAPTER 3: PHENOLIC COMPOUND ACCUMULATION ANALYSIS AND GENOME WIDE ASSOCIATION STUDY IN DIVERSE MAIZE KERNELS

ABSTRACT

Maize is an essential cereal crop used in many industrial and agricultural settings. Our society is incredibly dependent on the crop and there are many abiotic and biotic stressors that continue to threaten production resulting in inability to meet the growing demand. Phenolic compounds are small chemical compounds distributed throughout plant tissues that are connected with increased nutritional benefits and stressor resistance for the plant. Studying the phenolic compound accumulation in maize kernels will provide key insight on the genetic architecture of the phenolic compound accumulation in maize plants. Thirty-four phenolic compounds were analyzed in maize kernel tissue in varieties from the Wisconsin Diversity Panel. Twenty-one of the phenolic compounds were detected with high to moderate confidence and GWAS were run on each of the phenotypes. In total, there were 170 significant SNPs associated with phenolic compound accumulation in kernels which allowed the extraction of 390 candidate genes. The research done will provide insightful information to help aid breeding decisions and pave the pathway for future research on the candidate genes controlling phenolic compound accumulation.

INTRODUCTION

Maize (*Zea mays*) is an essential cereal crop used in many industrial and agricultural settings. It is considered a staple crop in the human diet and for animal feed. Due to the popularity and human dependence on maize, the nutritional elements and phytochemical compound accumulated are becoming increasingly prevalent in research studies (Rouf Shah et al., 2016). The Wisconsin Diversity panel is a group of 942 inbred lines. This population includes diverse public, expired plant variety protection (exPVP), and germplasm enhancement of maize (GEM) derived varieties (Mazaheri et al., 2019). These varieties include stiff stalk (SS), non-stiff stalk (NSS), Iodent (IDT), sweet corn, popcorn, tropical inbreds and a few of the unselected inbreds from synthetic populations and landraces (Mazaheri et al., 2019). The Wisconsin Diversity panel encompasses a large assortment of phenotypes used to study numerous traits.

Phenolic compounds, also called phenolics, are chemically identified as compounds that contain a hydroxylated aromatic ring, with the hydroxy group attached directly to the phenyl, substituted phenyl, or aryl group (Swanson, 2003). There are thousands of known phenolic compounds and they are distributed generously through plant tissues; phenolics are known to greatly contribute to plant immune response, defense, color, flavor, and astringency of plants (Swanson, 2003). Phenolic compounds can be classified into groups, including phenols, coumarins, lignins, lignans, condensed and hydrolysable tannins, phenolic acids, and flavonoids (Soto-Vaca et al., 2021). Flavonoids can then be separated into smaller subgroups based on the structural differences, these groups include chalcones, aurones, flavanones, flavones, dihydroflavonols, isoflavonoids, phlobaphenes, flavanols, flavonols, anthocyanins, and leucoanthocyanidins (Šamec, et al., 2021; Liu et al., 2021; Winkel-Shirley, 2021; and Chen et al.,

2023).

Flavonoids are highly abundant in plant tissues (Panche et al., 2016). Together flavonoids and phenolic acids make up the majority of dietary phenolic compounds and are found readily in the majority of fruits and vegetables, although, the type and concentration of flavonoids vary based on the plant variety and the tissue type (Xu et al., 2017; Erlund, 2004). Phenolic compounds, and specifically flavonoids, have crucial roles in plant metabolism and biology, leading to the biosynthesis pathways being a topic of interest to many researchers. The metabolism of phenolic compounds in plants include numerous different biosynthesis pathways within the plant (Lattanzio et al., 2012).

Flavonoids are known to be important components of plant metabolism, defense, and immune response, especially in cereals (Liu et al., 2013). Additionally, flavonoids are thought to protect cereal crops against numerous biotic and abiotic stressors including UV protection, insect resistance, disease resistance, developmental functions, pigmentation, and auxin regulation (summarized in Liu et al., 2013). Flavonoid biosynthesis and the intermediate metabolites created are essential to different processes that occur throughout the entire lifecycle of the plant. Phenolic compounds, mainly flavonoids, are attributed to seed dormancy, hormone regulation, root nodule development, and bacterial signaling (summarized in Liu et al., 2013). Flavonoids are also known to absorb UV rays which provides the plant with UV radiation protection needed for response to stress caused by excess UV rays (Bashandy et al. 2009).

In addition to the essential roles phenolic compounds play in plant metabolism, they also are highly nutritious and provide health benefits when consumed. Phenolic compounds, including flavonoids, are readily found in fruits and vegetables, and are thought to contribute to the many health benefits from a diet rich in fruit and vegetables (Ballard & Junior, 2019).

Although the mechanisms of flavonoids are of interest to many researchers, some of them are still not fully understood. The flavonoid mechanisms, either individually or in combination, have showed antioxidant, anti-inflammatory, anti-diabetic, anti-cancer, anti-obesity, and cardioprotective effects (Xiao et al., 2011; reviewed in Ballard & Junior, 2019). Previous research has been done on a small subset of flavonoids in specific maize tissues. The phenolic compounds include eriodictyol, luteolin, isoorientin, and maysin in the tissues of pollen, silks, tassels, and kernels (Zhang et al., 2018). This research found that there were variable amounts of phenolics in each of the different tissue types (Zhang et al., 2018). More research is needed on a wide variety of phenolic compound to understand the genetic control and architecture of phenolic compound accumulation in maize kernels.

A greater understanding of the phenolic compounds accumulated in maize kernels will provide a better understanding of the genetic architecture that controls the biosynthesis pathways and will create tools for researchers and breeders to implement in their maize breeding programs. A simplified scheme of phenolic compound biosynthesis and the pathways involved can be seen in figure 3.1, along with the phenolic compounds analyzed in this research study included in their respective categories. In this study, kernels from the Wisconsin Diversity Panel will undergo extraction and quantification procedures to measure the phenolic compound accumulation. These results will then be used for analysis on the phenolic compound distribution in maize kernels, to uncover significant SNPs through genome wide association studies (GWAS), and provide regions of the maize genome for future research of genes that control the accumulation of these compounds. The research in this study aims to explore the natural variation of phenolic compounds in diverse maize kernels, provide insight to plant breeders that want to capitalize on the benefits of phenolic compounds in their maize varieties, and highlight

potential candidate genes that help uncover the genetic architecture of phenolic compound biosynthesis.

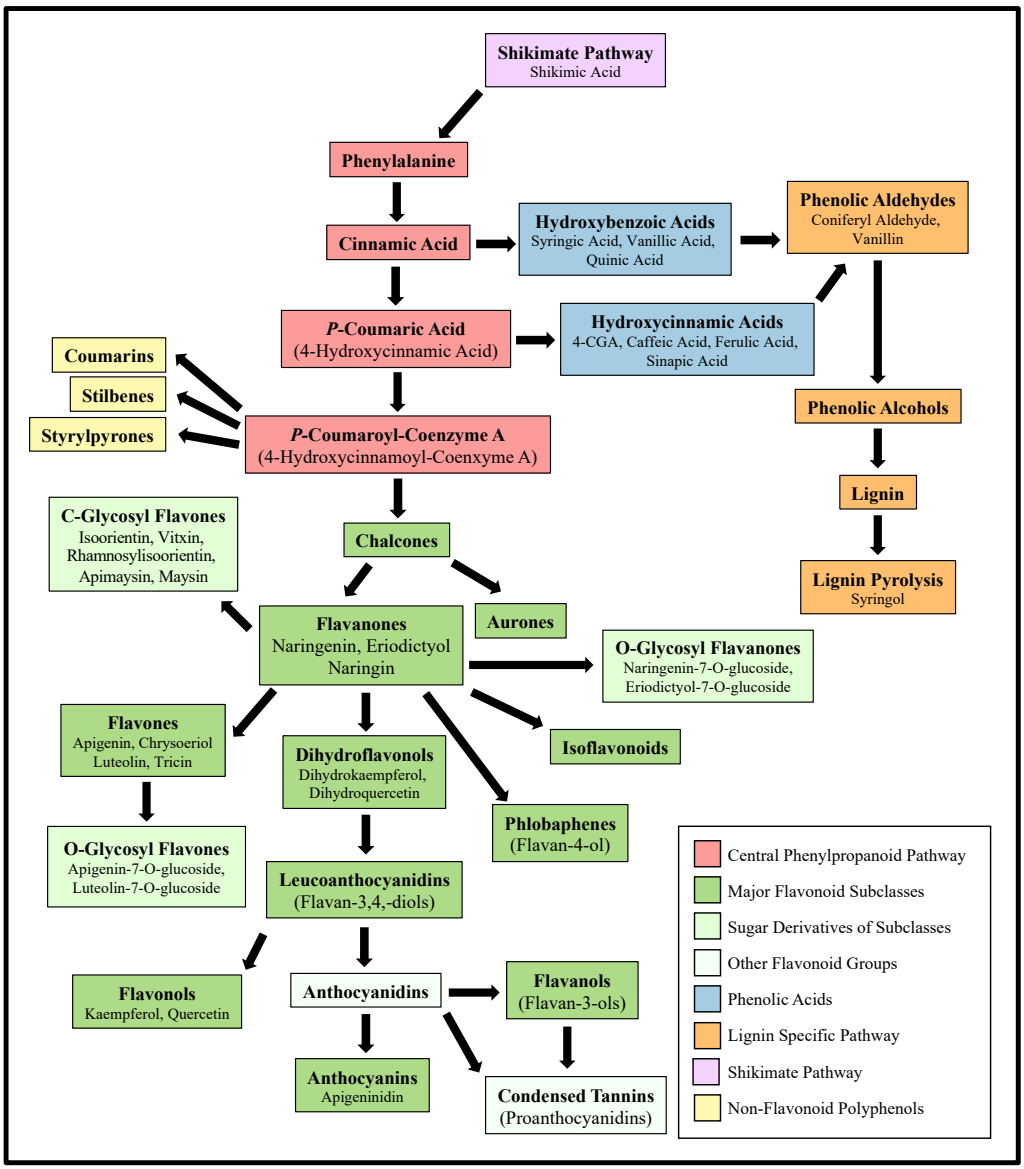


Figure 3.1: Generalized overview of the biosynthesis pathways involved in phenolic compound and flavonoid biosynthesis. Color indicates separate groupings of phenolic compound type or parts of the pathway. Arrows can either be describing a single step or can be simplifying multiple undescribed steps. Information in figure derived from Liu et al. (2021), Winkel-Shirley (2021), and Chen et al. (2023).

MATERIALS AND METHODS

Sourcing of plant materials and experimental design

The maize kernels utilized in this project were sourced from the germplasm resource information network (GRIN) global seed repositories. This experiment included 702 varieties that are all a part of the Wisconsin Diversity Panel. The kernels in this experiment when received and had appropriate moisture levels to indicate maturity upon harvest. A list of all the varieties used for the analysis can be seen in the supplemental files. Three kernels from each variety were used in the experiment as biological replicates. Each line was given an identification name including letters and numbers. To decipher between the kernels within a single variety, an underscore with the kernel number (1, 2, or 3) was added onto the line identification to become the sample identification name. Individual varieties were placed into labeled envelopes to maintain organization and reduce error.

Maize kernel lyophilization and grinding

Mature maize kernels were lyophilized for 5-7 days to complete a freeze-drying process and to ensure reduced moisture levels. Excess moisture in the kernels impacts the weight of true kernel tissue and skew analysis. After the kernels were adequately dried, they were cracked with either a pair of pliers or a hammer, to facilitate the grinding process, and then allocated into reinforced 2 mL bead mill tubes with three steel 5 mm beads. The kernels were then run through the Bead Mill tissue homogenizer. The tissue homogenizer was set at a speed of 5, grind time of 8 seconds, a dwelling time of 9 minutes and 59 seconds, and was run for 6 cycles. The kernels were then observed to verify that they were grinded into a fine powder. If any visible kernel chunks remained, they were resent through the bead mill homogenizer for 3 additional cycles on

the same settings. Once the kernel tissue was ground into a fine powder, 29.0-31.0 mg of kernel tissue was weighed and placed into 1.5 mL microcentrifuge tubes. Weight measurements were recorded to the nearest tenth of a milligram. The three kernels from each line were separated into individual tubes and considered independent samples or biological replicates.

Phenolic compound extraction via methanol solution

A solution of 80% MeOH and 0.1% formic acid was prepared and 500 µL was added to each sample. The sample tubes were then vortexed until the ground kernel tissue was completely resuspended. The samples were then stored in a 4° C refrigerator, undisturbed for at least 12 hours. After storing the samples, they were vortexed again to resuspend, then centrifuged at 13000 rpm for 3 minutes. Carefully, 400 µL of the supernatant was collected and reallocated into a fresh tube and was then diluted with sterile water to obtain a 50% MeOH solution concentration. The samples were vortexed again for 4-5 seconds and stored in a 4° C refrigerator until use. If the clump of ground kernel tissue was disturbed before the supernatant could be extracted, the sample should be re-centrifuged and then can be extracted.

Once the samples were stored in the 4° C refrigerator, they may become cloudy, or have visible kernel tissue powder in the extractant. If this happens, they were vortexed and put back in the microcentrifuge and re spun at the same settings as above. As much of the supernatant is then removed and kept in another tube and stored back in the refrigerator until use. This may have to be done multiple times until there is no sediment in the extracted liquid. The samples should be transparent, and any solids left may cause issues with future analysis.

Preparation of kernel extracts, quality control samples, and standard calibration curve

The kernel extracts were separated into three different batches. The batches had 201, 344, and 182 maize varieties, respectively. For each batch a pool was created for a means of quality control within the batch. The pool of samples was created using 100 μL of 100 different kernel extracts selected randomly. The pool of samples is aimed to be representative of the sample population as a whole and have phenolic compound concentrations similar to that of the population.

In order for accurate LC-MS data acquisition, standards of the phenolic compounds are needed to run alongside the maize samples. The phenolic compound standards used to make the calibration curve were sourced from Sigma-Aldrich (Burlington, MA, USA) Cayman Chemical (Ann Arbor, MI, USA), Indofine chemical (Hillsborough, NJ, USA), and ChromaDex (Irvine, CA, USA) except apimaysin, maysin, and rhamnosylisoorientin, which were provided by Michael McMullen (USDA-ARS) and Maurice Snook (Iowa State University) facilitated through Dr, Erich Grotewold (MSU), information on the standards can be found in Rodriguez et al. (2022). The calibration curve was made up of 34 phenolic compound standards in a series of serial dilutions, in a methanol solution, including 3.9 nM, 15.625 nM, 31.25 nM, 62.5 nM, 125 nM, 250 nM, 500 nM, 750 nM, and 1000 nM. The calibration curve of standards was pipetted into amber vials and run before every plate. The complete list of standards used for the calibration curve can be seen in table 3.1.

Table 3.1: List of the phenolic compounds used in the standard calibration curve for the LC-MS data acquisition. This table also includes the names of the phenolic compounds that are used in any accompanying data files.

Phenolic Compound	Name in Data Files	CAS No.	Classification	Sub-Class
4-Caffeoylquinic Acid	X4CGA	905-99-7	Phenolic Acid	Hydroxycinnamic Acid
Apigeninidin	Apigeninidin	1151-98-0	Anthocyanidins	Deoxyanthocyanidins
Apigenin	Apigenin	520-36-5	Flavone	Flavone
Apigenin-7-O-glucoside	Apigenin7Oglu	2492-87-7	Flavone	O-Glycosyl Flavone
Apimaysin	Apimaysin	-	Flavone	C-Glycosyl Flavone
Caffeic Acid	CaffeicAcid	331-39-5	Phenolic Acid	Hydroxycinnamic Acid
Chrysoeriol	Chrysoeriol	491-71-4	Flavone	Flavone
Coniferyl Aldehyde	ConiferylAldehyde	458-36-6	Phenolic Aldehyde	Hydroxycinnamaldehyde
Dihydrokaempferol	Dihydrokaempferol	480-20-6	Dihydroflavonol	Dihydroflavonol
Dihydroquercetin	Dihydroquercetin	480-18-2	Dihydroflavonol	Dihydroflavonol
Eriodictyol	Eriodictyol	4049-38-1	Flavanone	Flavanone
Eriodictyol-7-O-glucoside	Eriodictyol7Oglu	38965-51-4	Flavanone	O-Glycosyl Flavanone
Ferulic Acid	FerulicAcid	1135-24-6	Phenolic Acid	Hydroxycinnamic Acid
Isoorientin	Isoorientin	4261-42-1	Flavone	C-Glycosyl Flavone
Kaempferol	Kaempferol	520-18-3	Flavonol	Flavonol
Luteolin	Luteolin	491-70-3	Flavone	Flavone
Luteolin-7-O-glucoside	Luteolin7Oglu	5373-11-5	Flavone	O-Glycosyl Flavone
Maysin	Maysin	-	Flavone	C-Glycosyl Flavone
Naringenin	Naringenin	67604-48-2	Flavanone	Flavanone
Naringin	Naringin	529-55-5	Flavanone	Flavanone
Naringenin-7-O-glucoside	Naringenin7Oglu	501-98-4	Flavanone	O-Glycosyl Flavanone
p-Coumaric Acid	pCoumaricAcid	63-91-2	Phenolic Acid	Hydroxycinnamic Acid
Phenylalanine	Phenylalanine	849061-97-8	Amino Acid	Aromatic Amino Acid
Quercetin	Quercetin	77-95-2	Flavonol	Flavonol
Quinic Acid	QuinicAcid	-	Phenolic Acid	Hydroxybenzoic Acid
Rhamnosylisoorientin	Rhamnosylisoorientin	138-59-0	Flavone	C-Glycosyl Flavone
Shikimic Acid	ShikimicAcid	530-59-6	Phenolic Acid	Central Metabolite
Sinapic Acid	SinapicAcid	530-57-4	Phenolic Acid	Hydroxycinnamic Acid
Syringic Acid	SyringicAcid	91-10-1	Phenolic Acid	Hydroxybenzoic Acid
Syringol	Syringol	520-32-1	Phenol	Methoxyphenol
Tricin	Tricin	121-34-6	Flavone	Flavone
Vanillic Acid	VanillicAcid	121-33-5	Phenolic Acid	Hydroxybenzoic Acid
Vanillin	Vanillin	3681-93-4	Phenolic Aldehyde	Benzaldehyde
Vitexin	Vitexin	491-70-3	Flavone	C-Glycosyl Flavone

LC-MS analytical chemistry

The instrument used for the LC-MS analytical chemistry analysis was a Waters ACQUITY TQD Tandem Quadrupole UPLC/MS/MS (Waters Corporation, Milford, MA, USA). The three batches of maize kernel extract samples were run through the instrument alongside a standard calibration curve, pool quality control samples, blanks, and an internal standard. 90 µL of each sample was pipetted into 96 well plates. An internal standard of 10 µL of 500 nM 8-

prenylnaringenin was added into each well of the plate to bring the internal standard concentration to 50 nM. In each plate, 12 of the wells were filled with the pooled solution, 6 of the wells were filled with a blank methanol solution, and 18 of the samples were duplicated as technical replicates to take up 36 wells. The 96 well plates were run after a set of amber vials comprised of the calibration curve. This was done before each of the 96 well plates for higher quality samples.

The TQD Tandem Quadrupole UPLC/MS/MS was run on a targeted 10-minute targeted multiple reaction monitoring method setting to detect and quantify the compounds against the standard calibration curve. The targeted multiple reactions monitoring methods file was modified from Rodriguez et al. (2022). After the LC-MS data acquisition, the Mass Lynx program, Target Lynx was used for peak integration of the resulting chromatograms (Waters Corporation, 2021). Phenolic compound retention time was determined using the targeted methods file and observation of the standards in the calibration curve. The retention time was used to quality check and adjust the integrated peaks for each compound for each sample, as needed. The R^2 value for the standard curve was used as a way to determine quality of compound detection and standard solution quality. Points from the standard calibration curve were excluded if there was an obvious error with the sample. If more than 2 points needed exclusion, or the R^2 value was still under 0.95, the phenolic compound was excluded and was considered to have too much error to accurately detect and quantify the phenolic compound accumulation.

LC-MS data preparation, normalization, and cleaning

Concentrations of specific compounds fell outside the range of concentrations in the standard calibration curve (greater than 1000nM) and varying quality results for the internal

standard resulted in relative values used for compound quantification, instead of absolute values. Relative values are derived from the area of the integrated peak in the chromatogram and standardized with the calibration curve. To combat noise and machine error, compounds within samples were categorized as detectable only if the area was greater than 3 times the mean value of the blanks. This threshold is called the limit of detection. Any sample that did not meet that threshold, was given a value of NA for the phenotype detection of that compound.

Normalization of the samples was conducted using a calculation of arbitrary units of area (AUA). AUA was calculated with each compound in each sample by dividing the area of the integrated peak by milligram of kernel tissue per mL of methanol solution. Additional information on calculation of AUA can be seen in Rodriguez et al. (2022). This calculation accounted for the variation in weight of the ground kernel tissue during the weighing sample preparation after the tissue grinding.

The kernel samples were run through the TQD Tandem Quadrupole UPLC/MS/MS in three different batches with different pool samples. There was not a large overlap in the samples, so there was not a way to use the samples to account for the differences between batches. To help account for batch effect and to normalize the samples between the batches, the R package batchma was utilized (Stopsack et al., 2021). This package is used primarily for accounting and providing batch effect adjustments for biological phenotypes and biomarker data (Stopsack et al., 2021). Both the simple means and quantile normalization methods were tested, and the quantile normalization method (‘quantnorm’) was selected. Selection was done based on the highest heritability between the methods.

When using this type of data and excluding points based on limits of detection, some of the phenolic compounds had numerous missing points. Many missing values reduce the

confidence that the phenolic compound was accurately detected. Of the 34 phenolic compounds analyzed, 13 of them were detected with high confidence and had greater than 75% of the tested kernel varieties with detectable accumulation. An additional eight of the compounds were detected with moderate confidence, determined by greater than 200 of the varieties having detectable accumulation of the compounds. Lastly, there were 13 compounds that were not detected in more than 200 of the maize varieties, leading them to have low confidence and falling vastly outside the limits of detection. Table 3.2 includes a list of the phenolic compounds and whether they have low, moderate, or high confidence in their detection based on the above qualifications.

Table 3.2: The detection confidence for each of the phenolic compounds based on the limits of detection and the data cleaning processes. Color of detection confidence is a visual indicator of the confidence level of the detection of the phenolic compounds. Red is low confidence; yellow is moderate confidence; and green is high confidence.

Phenolic Compound	Detection Confidence
Apigeninidin	Low
Apigenin	High
Apigenin-7-O-glucoside	Moderate
Apimaysin	Low
Caffeic acid	High
Chrysoeriol	High
Coniferyl aldehyde	High
Dihydrokaempferol	Moderate
Dihydroquercetin	Moderate
Eriodictyol	Moderate
Eriodictyol-7-O-glucoside	Low
Ferulic acid	Low
Isoorientin	Low
Kaempferol	Low
Luteolin	Moderate
Luteolin-7-O-glucoside	Low
Maysin	Low
Naringenin	Moderate
Naringin	Low
Naringenin-7-O-glucoside	Low
p-Coumaric acid	High
Phenylalanine	High
Quercetin	Low
Quinic acid	High
Rhamnosylisoorientin	Moderate
Shikimic acid	Low
Sinapic acid	High
Syringic acid	High
Syringol	Low
Tricin	High
Vanillic acid	High
Vanillin	High
Vitexin	High
4-Caffeoylquinic acid	High

Phenolic compound distribution analysis

To understand the patterns of phenolic compound accumulation in diverse maize lines, further research into the mean accumulation for each compound compared with each other. To complete this distribution analysis, a mixed model using the r package lme4 was done to predict the total phenolic compound accumulation separately for each compound (Bates et al., 2015).

The equation used to fit the linear mixed model was $y_i = \mu + \alpha_i + e_i$. Where y_i is the phenolic compound accumulation of the i^{th} genotype, μ is the estimate for the mean of the phenolic compound, or the intercept, α_i is the genotypic effect for the i^{th} genotype with coefficient α, and e_i is the residual error for the i^{th} genotype. This model was used to calculate the estimate of the phenolic compound accumulation across all lines while accounting for the differences in genotypes with the random effects. The mixed model was calculated with the R package lme4, where the mean estimates, standard error, and *t* values were extracted (Bates et al., 2015). To understand the shape of the data and additional features on the phenolic compound distribution analysis, more summary statistics are calculated. Using the statistical software R, the mean, standard deviation, minimum, median, and maximum were calculated for each phenolic compound.

Phenotypic data analysis on maize variety subgrouping

Due to the varying background of the Wisconsin Diversity Panel, one of our research interests was to understand the distribution of phenolic compounds based on the varieties' subpopulation grouping. The subpopulation group were assigned based on origin and type of maize lines. The different allotted groups for subpopulation include broad origin-public, stiff stalk (SS), non-stiff stalk (NSS), Iodent (IDT), tropical, popcorn, sweet corn, and mixed. The SS group can be further divided into B73 or B37 derived and the NSS group can be separated into Mo17 and Oh43 derived, but for the analysis done in this study, only the general SS and NSS groupings were used. Information on the varieties' subgroupings were sourced from Mazaheri et al. (2019), Remington et al. (2001), Beckett (2016), CIMMYT, Mikel & Dudley (2006), GRIN global, Jarvis Golden Prolific, Liu et al. (2003), and De la Fuente (2015).

After the varieties were given a subpopulation grouping, the information was used as a factor as a means of predicting phenolic compound accumulation. A linear model was used with the formula $y_i = \mu + \beta_i + e_i$. Where, y_i is the phenolic compound accumulation of the i^{th} subpopulation group, μ is the estimate for the mean of the phenolic compound, or the intercept, β_i is the subpopulation group effect for the i^{th} subpopulation with coefficient β, and e_i is the residual error for the i^{th} subpopulation group. All effects in this model are considered fixed. Genotype was not taken into account for this model due to potential multicollinearity issues with genotype and subpopulation. An ANOVA was then performed for each of the models to determine if subpopulation grouping was significant when predicting the accumulation of the phenolic compound. Once it was determined there were significant differences between the subpopulation groups, the estimates were computed for the subpopulation groups in AUA units.

Genetic data preparation and Genome Wide Association Study

The marker files were sourced from Grzybowski et al. (2023) and were created by aligning whole genome resequencing data from 1,515 maize varieties to the maize B73 version 5 reference genome. The marker set was prepared by removing variants with alleles where more than 2 alleles were observed, variants with greater than 50% missing data, variants with either less than 1,515 or more than 33,550 sequence depth, and variants with inbreeding coefficients greater than 0 (Grzybowski et al., 2023). This created the starting marker file with ~46 million variants with imputation done with Beagle 5.0 (Browning et al., 2021). More information on the sequencing and alignment can be found in Grzybowski et al. (2023).

Using tassel version 5.0, each chromosomes file was transformed into HapMap format for ease in future analysis with R (Bradbury et al., 2007). The marker files were then read into R

and filtered based on the genotypes in common with the phenotyping file. There were 607 varieties in common between the genotype and phenotype files. The marker files were then read back into tassel version 5.0 and filtered for a minor allele frequency of 0.025 and a minimum count of 500 or greater number of taxa to be scored. After filtering, the chromosome marker files were uploaded onto Michigan State University's high performing computing cluster for greater computing power. The individual chromosome files were merged into a single genotype file in HapMap format through R, where all future analysis was conducted.

The genome wide association study (GWAS) to uncover significant SNP markers for our phenolic compound accumulation was conducted by GAPIT (version 3) using FarmCPU model (Wang & Zhang, 2021). Each phenolic compound phenotype data had a corresponding GWAS with 3 principal components. Numerous models and principal component combinations were sampled before final selection. Model selection and specifications were selected by observing QQ-plots resulting from the models. This work was supported in part through computational resources provided by the Institute of Cyber-Enabled Research at Michigan State University.

Extracting candidate genes from significant regions

There were numerous significant SNPs and GWAS hits resulting from the models run above. To distinguish the GWAS peaks from each other and decipher if there was any overlap between the significant SNPs, a linkage disequilibrium (LD) calculation was used. To complete the LD calculations, the genotype file was subset into only the rows that contained the significant SNPs and were loaded into tassel version 5.0 (Bradbury et al., 2007). Using the built in diversity feature, linkage disequilibrium was calculated between each combination of SNPs, this was done separately for each individual chromosome file. LD was calculated in the full matrix setting with

only the inbred genotypes included. The R^2 value was used to represent LD. The GWAS peaks were combined if they had an LD greater than 0.5 or were less than 50,000 bp apart. SNPs with LD of less than 0.5 were considered individual peaks (Mural et al., 2022).

To prepare the GWAS hits for the candidate gene extraction, all of the significant SNP results were compiled into a single file. Additional columns were added to the file to create a range for searching the genome 50kb less than and 50kb greater than the peak position. The B73 v5 reference genome (www.maizegdb.org) was used and transformed into a bed file for ease in analysis and extraction. Using the BEDTools function 'intersect', the intersection between the B73 v5 reference genome and the significant GWAS peak ranges were extracted along with the candidate genes (Quinlan & Hall, 2010).

Calculating genetic correlations among the phenolic compound phenotypes

To calculate the genetic correlations among the phenolic compound phenotypes, all of the SNPs that resulted in a significant GWAS hit were selected. This was not including any phenolic compounds that had a low detection confidence. This resulted in 170 SNPs that were deemed significant from the GWAS hits. Next, using the output from GWAS conducted by GAPIT (version 3) using FarmCPU model, the SNPs were ordered by the absolute value of the largest effect and the top 50 SNPs for each phenolic compound's GWAS were selected to create a list of SNPs that should represent the variety of phenolic compounds (Wang & Zhang, 2021). Any duplicate SNPs were removed, to only keep a list of unique SNPs IDs, this included 1137 SNPs. The genotype file used for the genetic analysis was then subset to only include the 1137 greatest effect SNPs. Using tassel version 5.0 the subset genotype file was formatted into numeric and output into a table format (Bradbury et al., 2007). Next, the numeric genotype file was

transformed to follow the rrBLUP format with genotype encoding of {-1,0,1} instead of the {0,0.5,1} format tassel uses. After obtaining the correct file formats, the 'mixed.solve' function, a part of the rrBLUP package in R, was used to calculate and extract the marker effects of the selected SNPs for each of the phenolic compound phenotypes (Endelman, 2011). Then the marker file was multiplied by the marker effects. Lastly, a Pearson correlation was calculated between the different phenolic compounds' SNP effects as a means of calculating genetic information. Heatmaps, made with the R package ggplot2, of the Pearson correlations between the phenolic compounds were used to visualize the genetic correlations between the phenolic compounds (Wickham, 2016).

Computing and visualizing genetic effects based on subpopulation of maize varieties

The subset genotype marker file of 1,137 SNPs and the marker effects previously calculated were multiplied together to prepare the data file for heatmap visualization. For each chromosome, the marker effects were added together to give a single value. This resulted in each phenolic compound and maize variety combination having ten values assigned to them, one for each chromosome. Then, using the ComplexHeatmap package in R, created by Gu (2016; 2022), the genotype effects were visualized separated by phenolic compound, subpopulation grouping, and chromosome. Color is used to indicate a positive effect or a negative effect on the accumulation of the phenolic compound in AUA.

RESULTS

Phenolic compound distribution for the Wisconsin Diversity Panel varieties

Of the 34 phenolic compounds analyzed, 13 of them were detected with high confidence

and had greater than 75% of the tested kernel varieties having detectable accumulation of the compounds. An additional 8 of the compounds were detected with moderate confidence, determined by greater than 200 of the varieties having detectable accumulation of the compounds. Lastly, there were 13 compounds that were not detected in greater than 200 of the maize varieties, leading them to have low confidence and falling vastly outside the limits of detection. The list of compounds and their detection confidence can be seen in table 3.2, with color indicating either low (red), moderate (yellow), or high (green) confidence of detection.

Phenolic compound accumulation summary statistics can be seen in table 3.3. The summary statistics explored are the mean, standard deviation, minimum, median, and maximum. These summary statistics were chosen to give an overall view of the spread of the data. Based on AUA, coniferyl aldehyde had the lowest mean estimate with a value of 0.17 AUA, and vanillin had the highest mean estimate with an accumulation of 442.78 AUA. There is a wide range in the estimates of the analyzed phenolic compounds in maize kernel tissue.

Table 3.3: Summary statistics for phenolic compound accumulation in maize kernels. Summary statistics include the mean estimate, standard deviation, minimum, median, and maximum for each compound across all samples.

Compound	Mean Estimate	Stdev	Min	Median	Max
X4CGA	17.03	60.01	1.51	3.82	781.09
Apigenin	1.98	2.99	0.22	0.81	26.45
Apigenin7Oglu	3.03	6.54	0.16	0.72	47.48
CaffeicAcid	14.18	16.90	0.87	8.49	184.95
Chrysoeriol	4.57	7.90	0.53	2.03	97.01
Dihydrokaempferol	0.33	0.66	0.05	0.11	5.45
Dihydroquercetin	0.24	0.46	0.04	0.08	3.09
Eriodictyol	0.65	1.33	0.11	0.30	18.68
Luteolin	0.26	0.50	0.06	0.12	4.54
Naringenin	6.27	7.78	1.75	3.59	57.09
pCoumaricAcid	209.77	114.40	30.78	184.21	872.11
QuinicAcid	9.91	6.64	2.69	7.66	45.87
Rhamnosylisoorientin	1.22	1.50	0.35	0.66	9.78
SinapicAcid	0.42	0.41	0.03	0.30	4.32
SyringicAcid	32.56	16.00	6.02	28.65	131.11
Tricin	14.10	23.00	1.18	6.25	224.70
VanillicAcid	169.51	72.07	43.20	159.02	578.15
Vanillin	442.78	164.31	124.83	416.74	1178.77
Vitexin	0.91	2.20	0.07	0.31	23.07
Phenylalanine	35.85	23.96	16.72	27.93	171.41
ConiferylAldehyde	0.17	0.16	0.03	0.12	1.32

Phenotypic analysis of phenolic compound distribution in subpopulation grouping

Each of the varieties sampled were then placed into categories that described their subpopulation or heterotic group. The 8 possible groups were SS (stiff stalk), NSS (non-stiff stalk), IDT (iodent), broad origin-public, sweet corn, popcorn, tropical, and mixed (Mazaheri et al., 2019). More in depth information on the population grouping structure can be found in Mazaheri et al. (2019), Remington et al. (2001), Beckett (2016), CIMMYT, Mikel & Dudley (2006), GRIN global, Jarvis Golden Prolific, Liu et al. (2003), and De la Fuente (2015). For this research 607 maize varieties were analyzed, of those, 138 are SS, 81 are NSS, 48 are IDT, 122 are broad origin-public, 25 are sweet corn, 15 are popcorn, 35 are tropical, and 143 are mixed. Linear modeling, with R, was used to predict the AUA value for each compound while taking in the fixed effect of subpopulation genetic group.

For all of the phenolic compounds except three, dihydrokaempferol, dihydroquercetin, and luteolin, subpopulation group was considered statistically significant when estimating accumulation of the phenolic compound, with a *p*-value threshold of 0.05. When looking at the phenolic acid classified phenolic compounds, including 4-chlorogenic acid, caffeic acid, p-coumaric acid, quinic acid, sinapic acid, syringic acid, and vanillic acid, subpopulation grouping is significant in all compounds. The sweet corn, tropical and IDT groups had the highest accumulation, on average, of the phenolic acid classified compounds, while the popcorn and SS group had on average the lowest accumulations of the phenolic acids. The flavones, including apigenin, apigenin 7-O-glucoside, chrysoeriol, luteolin, tricin, vitexin, and rhamnosylisoorientin, showed different results than the phenolic acids. Luteolin showed no significant differences between the subpopulations. The NSS subpopulation had the greatest accumulation of 4 of the other 6 flavones, then followed closely by broad origin-public with the second highest accumulation, while the sweet corn group had the lowest accumulation for 3 out of the other 6 flavones, with tropical having, on average, the second lowest means. The flavanones, eriodictyol and naringenin, did not show any clear subpopulation groups' accumulations in common. The flavonols, dihydrokaempferol and dihydroquercetin, both showed no significant statistical differences between the different genetic subpopulations. The phenolic aldehydes, vanillin and coniferyl aldehyde, had the highest accumulation in the tropical subpopulation and the lowest accumulation in the SS. The other group includes phenylalanine only, which is an aromatic amino acid, shows the greatest accumulation in the IDT and NSS groups, and the lowest accumulation in the SS and broad origin-public group. A color-coded table of all the phenolic compound accumulations distributed over the subpopulations can be seen in table 3.4, with the two highest and two lowest accumulated subgroups highlighted.

Table 3.4: Phenolic compound accumulation among the different subpopulation genetic groups, clustered by subclassifications. Color indicates the top two highest means and the top two lowest means. Dark red is the highest mean, light red is the second highest mean, dark blue is the lowest mean, light blue is the second lowest mean. Rows where the means are grayed out indicate no statistically significant differences between the subpopulation groups.

Phenolic Compound	Subpopulation Genetic Group							
	SS	NSS	IDT	Broad-Origin Public	Tropical	Sweet Corn	Popcorn	Mixed
Phenolic Acids								
4-Chlorogenic Acid	27.28	11.74	6.61	16.66	21.74	8.8	3.99	16.52
Caffeic Acid	12.1	14.3	18.3	12.8	16.3	24.8	18.8	15.1
pCoumaric Acid	184.21	236.04	247.98	200.80	262.81	213.61	237.10	207.32
Quinic Acid	10.29	10.39	8.84	9.73	9.44	13.39	6.46	10.02
Sinapic Acid	0.37	0.44	0.58	0.39	0.55	0.56	0.38	0.40
Syringic Acid	30.5	32.5	36.2	32.5	40.6	29.0	24.9	34.5
Vanillic Acid	163.5	134.7	192.5	176.98	168.9	187.4	205.5	184.2
Flavones								
Apigenin	1.82	2.77	1.06	2.16	1.05	0.80	1.47	1.93
Apigenin 7-O-glucoside	1.664	5.191	1.157	3.638	0.484	0.287	0.819	2.739
Chrysoeriol	3.73	4.36	4.02	7.80	1.87	1.55	2.47	4.63
Luteolin	0.196	0.205	0.165	0.329	0.305	0.144	0.355	0.21
Tricin	13.4	21.6	18.4	12.5	5.2	12.6	9.2	15.1
Vitexin	0.482	1.895	0.295	0.831	0.815	0.781	0.771	0.811
Rhamnosylisoorientin	1.868	1.225	1.205	1.166	0.516	0.851	0.587	1.045
Flavanones								
Eriodictyol	0.368	0.76	0.462	0.581	0.748	0.599	0.348	0.901
Naringenin	6.83	8.82	4.91	5.3	7.59	4.22	9.06	5.25
Dihydroflavonols								
Dihydrokaempferol	0.344	0.412	0.097	0.296	0.475	0.598	0.552	0.268
Dihydroquercetin	0.208	0.209	0.136	0.274	0.162	0.301	0.403	0.294
Phenolic Aldehydes								
Coniferyl Aldehyde	0.125	0.236	0.176	0.149	0.255	0.235	0.241	0.158
Vanillin	402.97	437.5	510.5	412.97	517.7	484.8	516.83	441.4
Other								
Phenylalanine	30.4	42.8	44.4	33.6	39.1	37.4	35.3	37.7

Correlation and relationships between phenolic compound accumulation

The phenolic compounds analyzed in this experiment are all a part of complex biosynthesis pathways resulting in different correlations and patterns relating to their production, use, and accumulation. Using a Pearson correlation and the R package ggplot2 for analysis and visualization of the correlations, we are able to see groupings of compounds that have positive correlations, or slightly negative correlations (Wickham, 2016). A heatmap of the phenolic

compound correlations among each other can be seen in figure 3.2, with the red color indicating a positive correlation, and a blue color indicating a negative correlation. The largest positive correlations are between apigenin 7-*O*-glu and apigenin, with a value of 0.62, apigenin 7-*O*-glu and chrysoeriol, with a value of 0.53, apigenin and chrysoeriol, with a value of 0.51, and syringic acid and p-coumaric acid, with a value of 0.5. The lowest correlation was a value of -0.13 and if was found between both dihydroquercetin with rhamnosylisoorientin and luteolin with syringic acid. The same results were seen when phenolic compound AUA values were averaged over the subgroups and then compared for correlations.

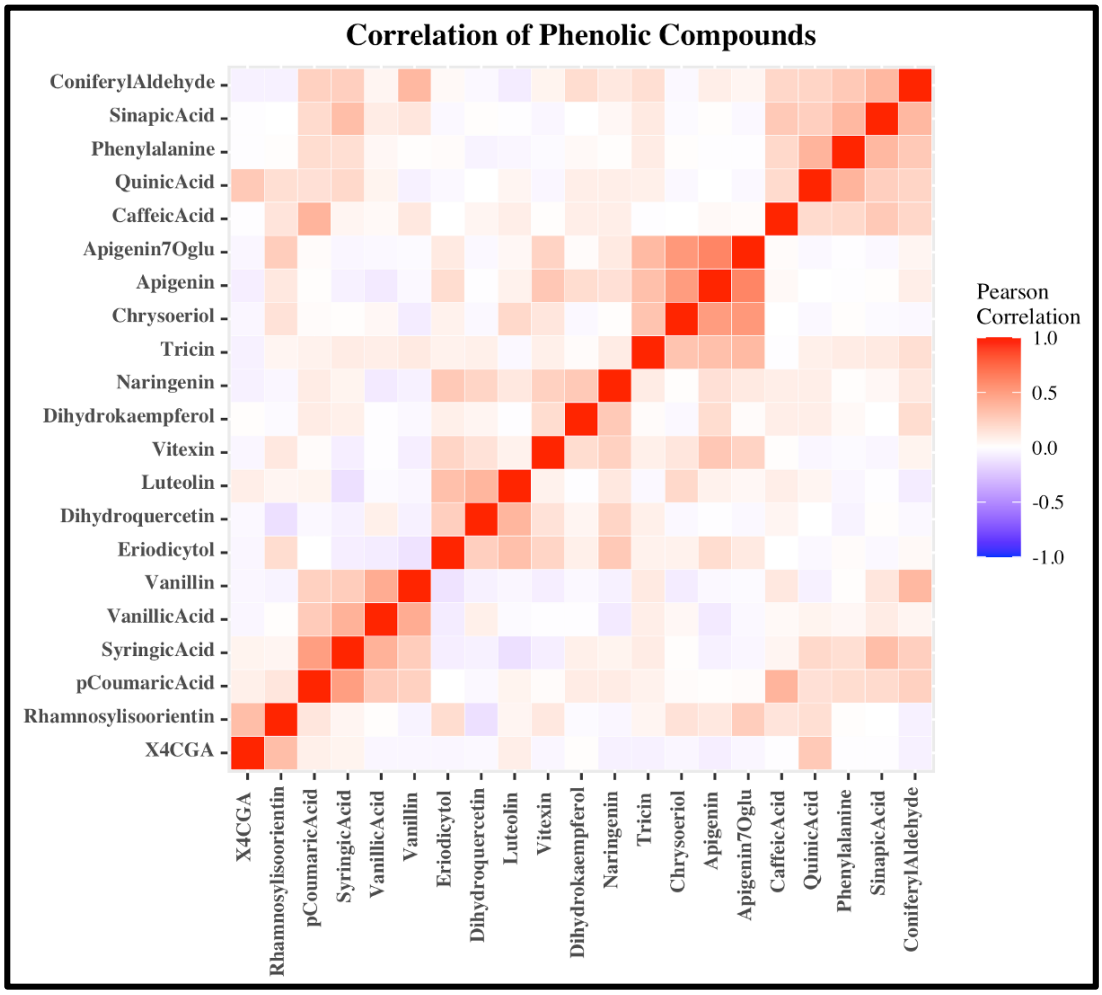


Figure 3.2: Heat map of Pearson's correlations between phenolic compound accumulation in diverse maize lines. Color indicates the strength and direction of the correlational relationship. Red indicates a positive correlation and blue indicates a negative correlation. Figure created with the R package ggplot2 (Wickham, 2016).

Genome wide association study and resulting significant SNPs

Each phenolic compound phenotype was run on a separate GWAS. The GWAS was run with GAPIT (version 3) using FarmCPU model with 3 principal components (Wang & Zhang, 2021). Model selection was made based on the QQ plots. A successful QQ plot was determined to have a large selection of the SNPs close to the line and only the end of the line showing points that converge to show greater significance. An example of two of the QQ plots can be seen in figure 3.3. The rest of the QQ plots can be found in the supplemental files.

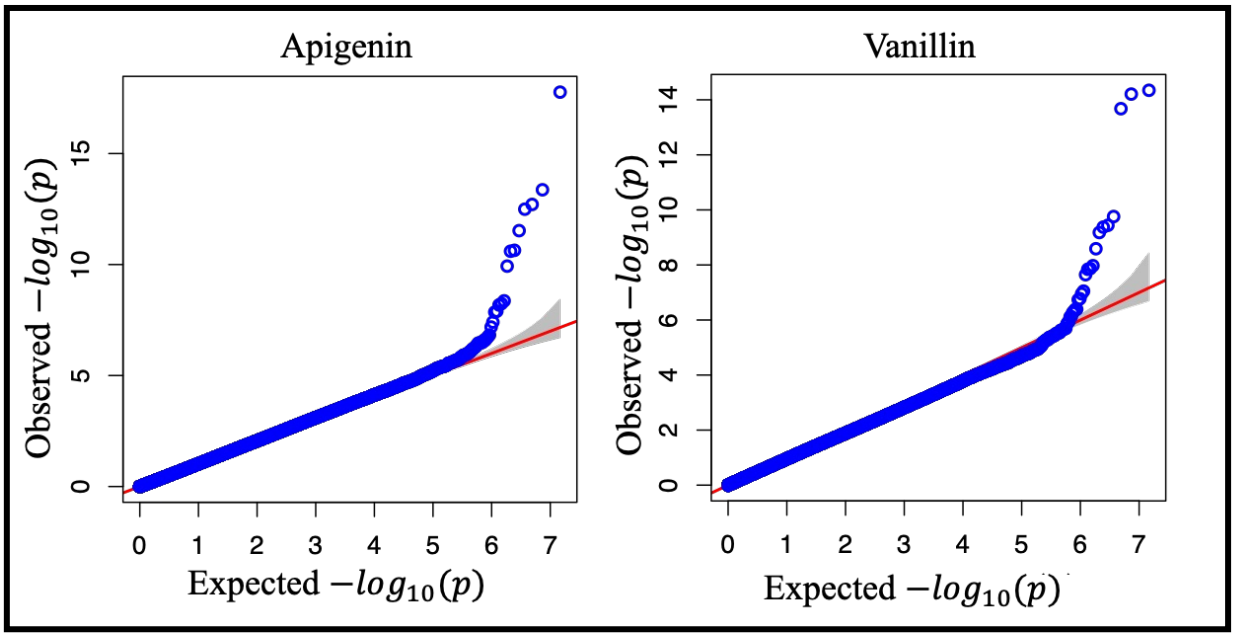


Figure 3.3: QQ-plots for the phenolic compounds, apigenin and vanillin. The QQ-plots were created using GAPIT version 3 with the FarmCPU model and 3 principal components (Wang & Zhang, 2021). Points on the plots that stray far from the red line indicate significant SNPs.

In total, for the 21 different GWAS run on each of the compounds, 19 of the 21 resulted in significant GWAS hits. The two phenolic compounds that did not result in a significant GWAS hit were p-coumaric acid and quinic acid. For the 19 other compounds, there was 170 significant SNPs highlighted, spanning across all 10 chromosomes. A complete list of the SNPs can be found in the supplemental files. Combined Manhattan plots with only the significant SNPs can be seen in figure 3.4. Chromosome 5 had the most significant SNPS with a total of 28, including GWAS hits from 14 different phenolic compounds. While chromosome 6 only had 8 GWAS hits from 7 different phenolic compounds. The phenolic compound vitexin had the most GWAS hits with 17 across all 10 chromosomes.

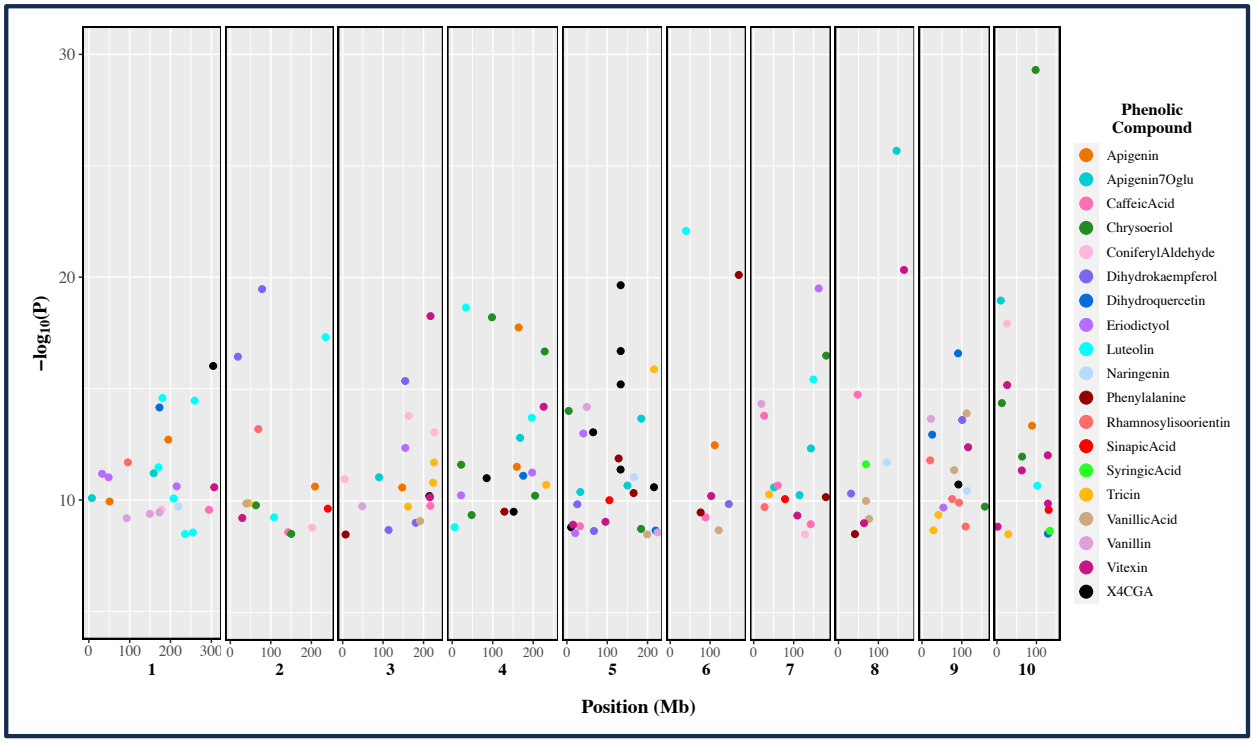


Figure 3.4: Combined Manhattan plots for significant GWAS hits for the phenolic compound accumulation analysis. 19 of the 21 phenolic compounds resulted in significant GWAS hits, which totals 170 highlighted SNPs spanning across all 10 chromosomes. Plot created with the R package ggplot2 (Wickham, 2016). Color indicates the different phenolic compounds and all points on the graph are from significant SNPs highlighted by the GWAS.

Candidate genes extracted from GWAS peaks

Separating the GWAS peaks is essential for extracting the candidate genes from the correct regions and not overrepresenting a single GWAS peak. Three of the peak locations had multiple SNPs that needed to be merged due to the lack of distance between them and the high LD. The first was on chromosome 1, around position 172,979,560 bp from the phenolic compounds dihydroquercetin and vanillin. There were two SNPs that were 33,266 bp apart and combined to a single peak taking the SNP with the highest *p* value as the main SNP. On chromosome 5, there were 4 SNPs around the location 133,293,673 bp all from the phenolic compound X4CGA. These four SNPs were merged into a single peak due to the high LD R^2 values between them. Lastly, on chromosome 8, two significant SNPs 647 bp apart were merged into a single peak at

location 69028460 bp.

Candidate genes were extracted from 50,000 bp before and after the GWAS peak. In total, there were 390 candidate genes extracted from the GWAS peaks. Some of the significant SNPs did not have any genes within the surrounding range, these SNPs were left off of the supplementary table. The complete list of candidate genes can be seen in the supplemental files.

Genetic correlational relationships between phenolic compounds

Absolute value of SNP effect results from the FarmCPU model from GAPIT version 3 resulted in 1,137 unique SNPs used for the rrBLUP model. The largest positive correlation was between apigenin-7-O-glugoside and apigenin with a genetic correlation of 0.59. Both apigenin and apigenin-7-O-glucoside are flavones, the only difference is apigenin-7-O-glucoside is an O-glycosyl flavone and has a different sugar attachment. The second and third highest positive correlation are values of 0.47 and 0.46 between syringic acid with p-coumaric acid and phenylalanine with quinic acid, respectively. Syringic acid and p-coumaric acid are both phenolic acids. Phenylalanine and quinic acid are not classified as the same type of compound, with phenylalanine being an amino acid and quinic acid is a phenolic acid. The three largest negative correlations have values of -0.22, -0.21, and -0.20 between the pairs naringenin with vanillin, apigenin with vanillic acid, and vanillin with eriodictyol. All of the above pairs occur in phenolic compounds that do not fall into the same subclass. The genetic correlations between the compounds can be visualized in the heatmap in figure 3.5 and the data can be found in the supplemental files.

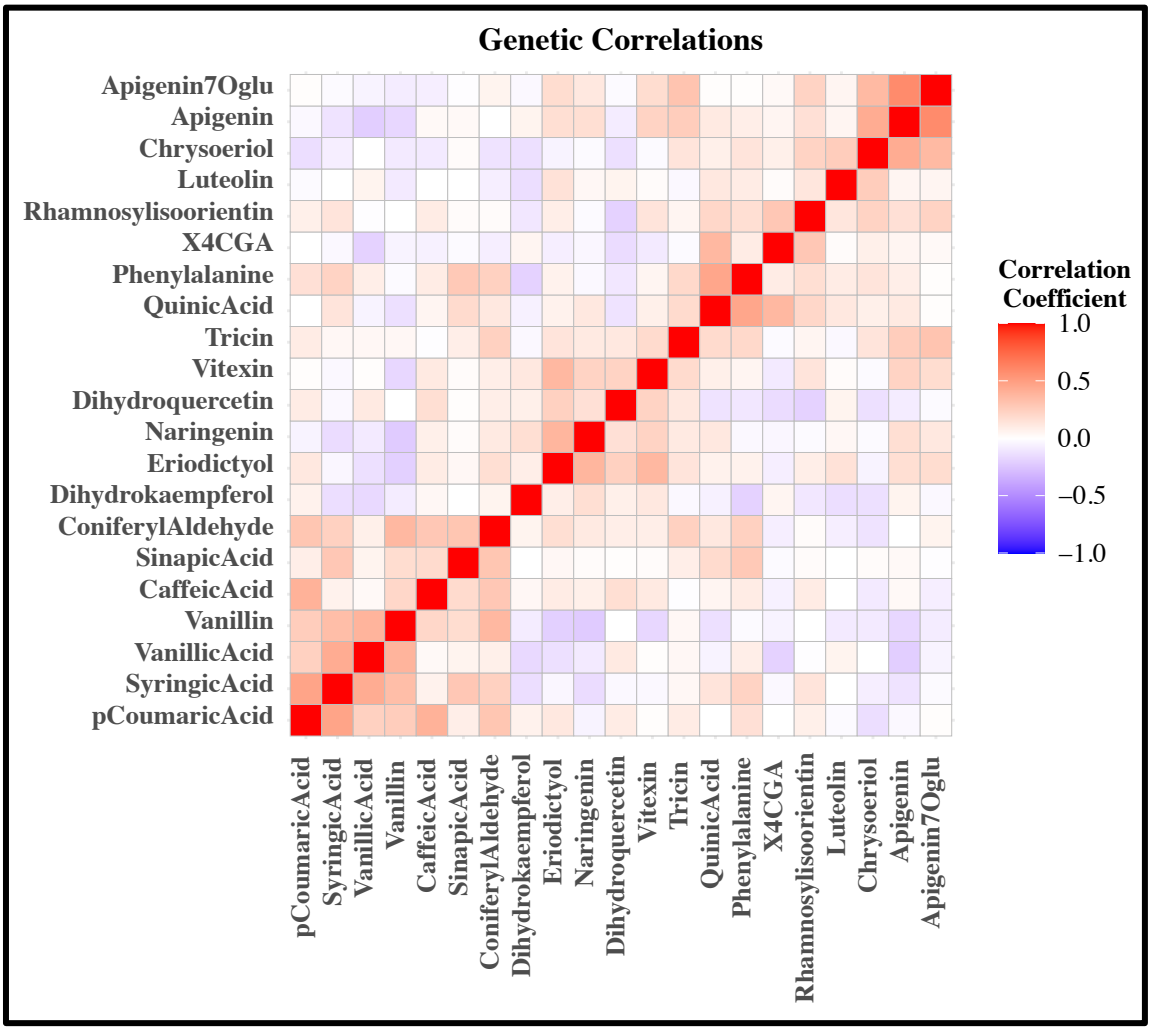


Figure 3.5: Heatmap of genetic correlations between the phenolic compounds. Figure created with the R package ggcorrplot (Kassambara, 2022). Color indicates strength and direction of correlational relationship with red indicating a positive correlation and blue indicating a negative correlation.

DISCUSSION

Phenolic compound distribution in maize kernels

Phenolic compounds have been prevalent in recent research due to their connections to plant immune response and nutrition. The phenolic compound accumulation in maize kernels has not been studied as widely as other tissues. Of the 34 phenolic compounds studied, 13 of them were not accumulated confidently in the varieties. It cannot be concluded that these compounds are exclusively not in maize kernels, but more so that they were not widely found in the varieties

examined during this study. The phenolic compounds may have been found above the limits of detection in a few of the varieties, but confident conclusions on the presence in maize kernels as a whole cannot be made on the limited number of samples. An additional 7 of the phenolic compounds had a moderate confidence in detection with at least 200 varieties surpassing the confidence of detection threshold. The final 14 compounds had high confidence in their overall detection in maize kernels with more than 75% of the varieties showing presence of the phenolic compound. These results cannot speak to maize kernels individually or other tissue varieties, only as a general view of the samples analyzed in this study. Other extraction procedures, or an adjusted targeted methods file for the TQD Tandem Quadrupole UPLC/MS/MS may lead to different results on the detection limitations.

Explanation of genetically diverging subpopulations

Domestication of maize years ago has led to numerous varieties and genetic subpopulations for what we call maize varieties today. Each of the maize subpopulations have diverging genetic backgrounds that are used to place them into groups. Domestication of maize began around 9000 years ago to form multiple variants of the single *Zea mays* species (Ngapo et al., 2021). During this time period, it was described that there were 5 different types of corn; these types include, soft, flint, sweet, pop, and pod (Parker, 1910).

Popcorn is a variant of flint corn that is popular globally. Flint corn was selected for their high sugar and starch content, low oil content, and a thick seed coat (Ngapo et al., 2021). Popcorn differs from other versions of flint corn due to the size and shape of the kernels and the ability for the kernels to burst when heated (Ziegler, 2003). Over the course of domestication and repeated selection for the color, size, shape, and popping quality of the flint corn, popcorn was

developed as a variant. Through the process of domestication and selection, many traits were selected for and selected against, this may have unknowingly resulted in significantly different phenolic compound accumulation of popcorn. Popcorn varieties, as a subpopulation in this experiment, show some of the lowest accumulation values for phenolic acids, low-moderate results for flavones, and some of the highest results for phenolic aldehydes. The significantly different phenolic compound accumulation results for this subpopulation could be an accompanying result to the numerous years of selection for desirable popcorn traits.

Sweet corn is thought to have two potential origins. The first theory is that modern sweet corn descended from Maiz Dulce and Chullpi, a sweet corn indigenous of Mexico and South America, respectively (Tracy, 2000). The second theory is that modern sweet corn results from a relatively recent mutation of *su1* allele in field corn (Tracy, 2000). Sweet corn differs from other varieties of field corn due to the flavor, aroma, and texture of the endosperm. Flavor of sweet corn is determined by the amounts of sugar and starch in the endosperm. Sweet corn breeding is centered on creating high sugar varieties through manipulation of the endosperm genes that control the levels of sugar and starch. Sweet corn shows some of the highest accumulations in phenolic acids and some of the lowest accumulation of flavones. This could be an indirect result from the increased selection of sugar content over the years.

Stiff stalk maize varieties are one of the main heterotic groups used for developing commercial maize hybrids. The stiff stalk heterotic group originated from the 1930s where G.F. Sprague developed a population from 16 inbred lines with increased stalk strength, called the Iowa Stiff Stalk Synthetic (BSSS) (Hallauer, 2009). This population was created to address the prominent issues with stalk lodging (Hallauer, 2009). This heterotic group has gone on to develop numerous successful inbred lines for hybrid development, while encouraging genetic

improvement. The stiff stalk inbreds analyzed in this research showed some of the lowest accumulation of phenolic acids, phenolic aldehydes, and phenylalanine. The Non-Stiff Stalk heterotic group is another one of the main groups of inbreds popularly used to make commercial maize hybrids. This population was created to be extremely genetically diverse when compared to the Stiff Stalk heterotic group. The creation of these separate heterotic groups has helped maintain genetic diversity in maize for gains made during hybrid breeding (Lu & Bernardo, 2001). Due to the contrasting genetics and the breeding for separation between the groups, it would be expected that some of the genes controlling the phenolic compound biosynthesis and accumulation could be impacted between the two groups. The non-stiff stalk group showed on average the highest accumulation of flavones, one of the highest accumulation patterns of flavanones, second highest accumulation of phenylalanine. The accumulation patterns of the stiff stalks and non-stiff show contrasting patterns and the selection methods for diverging genetics to create elite hybrids could be part of the cause of the significant differences.

The Iodent maize heterotic group makes up the third of the main groups used for hybrid commercial breeding and make up a significant portion of modern maize germplasms. The Iodent maize heterotic group originated from an ear to row breeding program with the goal of producing early maturing lines (Barrière et al., 2006). Preliminary iodent lines were bred for high yields, disease tolerance, and resistance to European corn borer; the varieties were then selected for improvement of flowering date and drying rate for the creation of early maturing varieties (Barrière et al., 2006). As one of the three main maize heterotic groups, Iodent varieties have been bred to be genetically diverse from the other two heterotic groups, it could be part of the cause of the different phenolic compound accumulation patterns. The Iodent group shows some of the highest accumulation patterns of the phenolic acids, and moderate accumulation of the

other phenolic compound subgroups.

Tropical subpopulation genetics of maize refers to varieties that are adapted to tropical environments. Tropical environments can cause additional stressors to the plants that they do not face in less harsh climates. Environments in the tropics can include extreme weather variations, intense rainfall, drastically changing temperatures, high pest and disease pressures, and low soil fertility (Pandey & Gardner, 1992). The tropical subpopulation group has on average one of the highest accumulations of phenolic acids, and one of the lowest accumulation patterns of flavones. The selection and breeding for the specific environmentally advantageous traits could provide an explanation for the phenolic compound accumulation results.

Determining how genotypic effects differ between the subpopulation groups

Maize varieties are placed into their subpopulation groups based on the genetic origin of the variety. Patterns emerge when looking at the genotype by marker effects when they are split into the heterotic groupings used to distinguish between the lines of the Wisconsin Diversity Panel. Heatmaps, made using Complex Heatmaps package in R, are used to visualize the genotypic effects divided by subpopulation grouping and chromosome (Gu, 2016 & 2022). The genotypic effects are visualized in figure 3.6 with red showing positive phenolic compound accumulation and blue resulting in negative phenolic compound accumulation. Genotypic effects were configured by multiplying the SNP effects found using rrBLUP modeling, buy the numeric marker matrix in the {-1, 0, 1} format. Using the heatmaps, we are able to broadly identify chromosomes within specific subpopulation groups that are responsible for higher or lower phenolic compound accumulation. For example, figure 3.6 shows that the phenolic compound vitexin has greater accumulation in chromosome 6 of the sweet corn subpopulation than the other

chromosomes. Another example is the large negative (blue) genotypic effects section on chromosome 6 for Tricin in the sweet corn group stands out when compared to the other groupings.

This heatmap can be used to distinguish important chromosomal regions from the subpopulations for higher or lower phenolic compound accumulation. Insight from the genotypic effects heat map can be used to isolate subpopulations or varieties that are significantly higher or lower than the others. Identifying specific subpopulations and chromosomes that have the genetic architecture that influences phenolic compound accumulation can lead to a better understanding of the biosynthesis pathways, the genes that control them, and potential breeding choices.

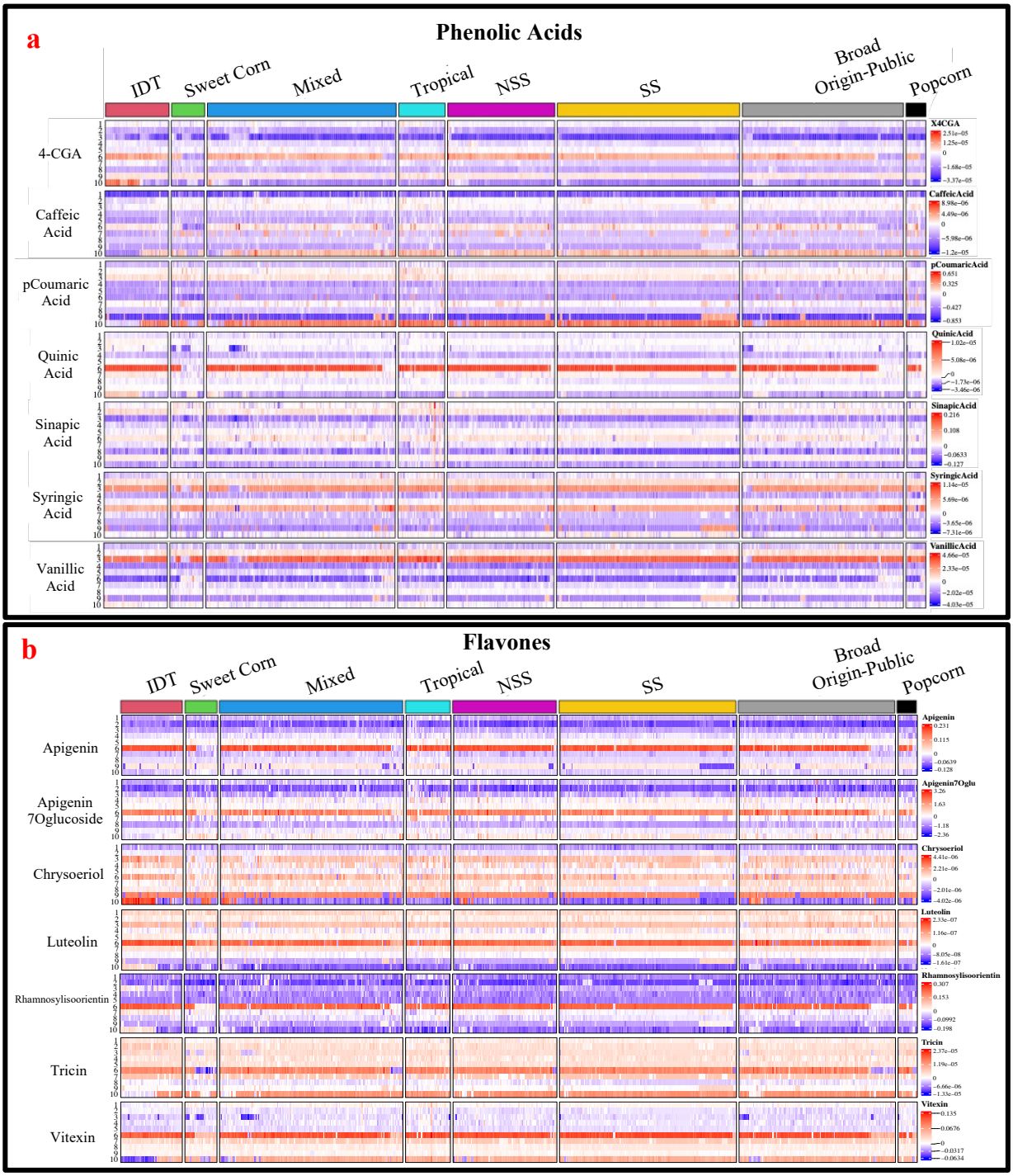
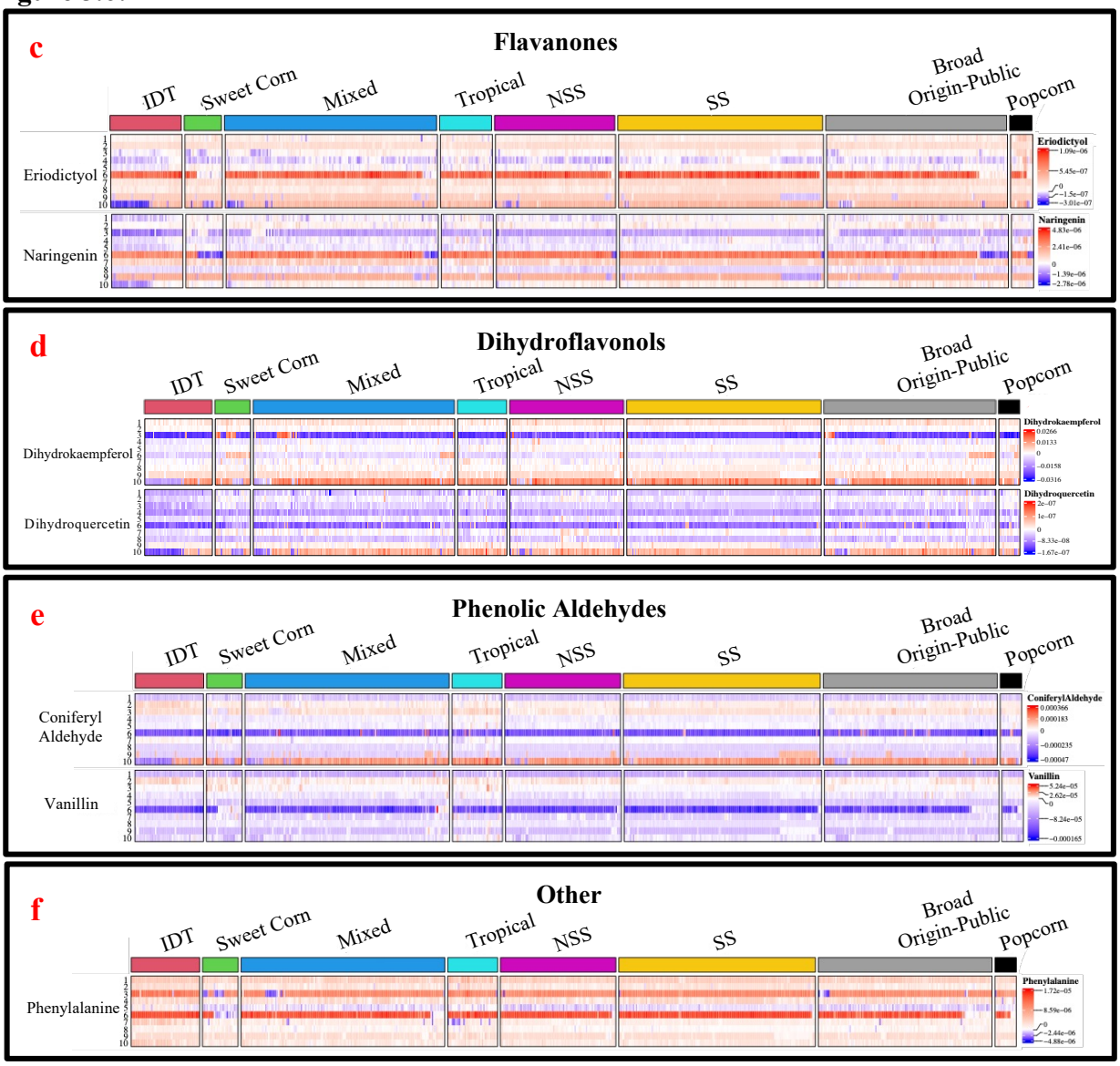


Figure 3.6: Genotypic effects by chromosome for the different maize subpopulation genetics groups, divided by phenolic compound subclassification. (a) – Phenolic acids, (b) – flavones, (c) - flavanones, (d) – flavanonols, (e) – phenolic aldehydes, (f) – other. Genotypic effects for 1137 SNPs across the genome, averaged by chromosome for each maize variety. The 10 rows for each phenolic compound indicate the 10 maize chromosomes. Red values indicate coloring indicates positive affect of the chromosome of a specific variety on the accumulation of the phenolic compound, and blue indicates the negative affect of the chromosome. Figure made with the R package, ComplexHeatmap (Gu 2016 & 2022).

Figure 3.6: Cont'd



Using phenolic compound accumulation data to aid in future breeding decisions

Phenolic compounds are known to have numerous desirable effects on the plant and those who consume them. Understanding the genetic architecture of phenolic compound accumulation in the kernels of maize plants gives researchers and plant breeders additional information on how to increase the accumulation of these nutritious elements within their developing varieties.

If breeders have interest in a specific phenolic compound or class of compounds, they can

use either the table of accumulation values broader information on the genotypic effects split by subpopulation group to introgress traits into their program. For example, flavanones are known to act as antioxidants and have been discovered to play a significant role in the anti-inflammatory response (Bredsdorff et al., 2010; Soto-Vaca et al., 2012). If a breeder intends on using the maize kernels for food or feed purposes and has interest in increasing the antioxidants and anti-inflammatory properties of the maize kernels, they may want to increase the quantity of flavanones, like naringenin and eriodictyol, in their varieties. To do this, the breeder may take insight from table 3.4 and figure 3.6 and see that the popcorn, mixed origin, and NSS subpopulation groups have the highest accumulation of these flavanones and chromosome 6 show the highest positive accumulation of these genotypic effects. This use of the information provided may allow the breeder to increase the accumulation of the flavanones.

Another example includes looking at the dihydroflavonols of dihydroquercetin and dihydrokaempferol. Looking at table 3.4, there is no significant differences between the accumulation in each of the subpopulation groupings, so this would not be a useful tool for breeding programs. Analyzing the map of figure XXX, we can see that overall, chromosome 10 has the highest positive effect on accumulation of the dihydroflavonols. It is also important to look at the genotypic effects of the first half of the IDT subpopulation grouping for chromosome 10. While most of chromosome 10 shows a positive overall effect for the dihydroflavonols, A large percentage of the IDT varieties show a negative effect on the overall accumulation. In addition to the effects of chromosome 10, chromosome 3 for dihydrokaempferol and chromosome 6 for dihydroquercetin have the largest negative effect on the accumulation in the kernel tissue. A breeder or future researchers would be able to investigate the effects of chromosome 10 for dihydroflavonols on the accumulation while trying to mitigate the genes on

chromosome 3 and 6 that limit the accumulation. The breeder may also avoid introgressing maize varieties with an IDT genetic background if they are trying to increase the accumulation of dihydroflavonols in the kernels.

The numerous benefits from phenolic compounds lead them to be a desirable phenotypic trait for breeding purposes. A greater understanding of the genetic architecture on the phenolic compound accumulation and the genes that are involved will allow this to be a more accessible tool for maize breeders. Future research on narrowing down the candidate genes discovered through this research will provide greater confidence in the function of the genes and how they affect phenolic compound accumulation specific to the maize kernel tissue. While the results discovered in this study are insightful and provide greater knowledge on the diversity of phenolic compound accumulation, more research is needed to validate the findings.

Progression of phenolic compound accumulations from maize kernels to maize seedlings

Using the results from Gomez Cano et al. (2023), we are able to compare the accumulation of phenolic compounds in maize kernel tissue and in maize seedlings. The seedling tissue samples were prepared in the same way as the kernels, as well as run on the same machine, the same methods file, and the same sample analysis procedures. 27 of the phenolic compounds were detected with high confidence, those that weren't detected are: apigenidin, caffeic acid, syringol, luteolin7Oglu, dihydrokaempferol, syringic acid, and naringin. The kernel and seedling data showed 18 phenolic compounds detected in common between the tissue types. To compare the accumulation of each phenolic compound by tissue type, a mixed effects model was used with the equation $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$. In this model, μ is the intercept, which is a fixed effect, α is the effect of the i^{th} tissue type of the sample, which is a fixed effect, β is the estimate

of the j^{th} genotype as a random effect, e is the residual error of the i^{th} tissue type of the j^{th} genotype, and y_{ij} is the accumulation of the phenolic compound of the i^{th} tissue type of the j^{th} genotype. The R packages lme4 and emmeans were used for the creation of the mixed model and to conduct an ANOVA (Bates et al., 2015; Lenth, 2023).

Of the 18 phenolic compounds in common, 16 of them showed significantly statistically different accumulation patterns. The only two that showed non-statistically significantly different patterns are 4-CGA and apigenin-7-O-glucoside, with *p*-values of 0.3462 and 0.5162, respectively. The other 16 phenolic compounds showed significant differences, all with p-values <0.0001. A table of the mean accumulations, fold changes, and p-values can be seen in table 3.5. Kernel tissue showed higher accumulation of the phenolic compounds p-coumaric acid, phenlalanine, quinic acid, tricin, vanillic acid, and vanillin. Seedling tissue showed higher accumulation of the phenolic compounds apigenin, chrysoeriol, coniferyl aldehyde, dihydroquercetin, eriodictyol, luteolin, naringenin, rhamnosylisoorientin, sinapic acid, and vitexin. Boxplots of the phenolic compound accumulation can be seen in figure 3.7.

Table 3.5: Phenolic compound accumulation comparisons between the maize kernels and seedlings. Maize seedling data from Gomez-Cano et al. (2023). The mean columns are measured in AUA and the fold changes are measured by the seedling mean divided by the kernel mean for each compound.

Phenolic Compound	Kernel Mean (AUA)	Seedling Mean (AUA)	Fold Change (Seedling/Kernel)	*p*-Value
4-CGA	17.03	14.18	0.83	0.3462
Apigeninidin	Not Well Detected	Not Well Detected	NA	NA
Apigenin	1.98	3.63	1.83	<0.0001
Apigenin-7-O-glucoside	3.03	3.02	1.00	0.5162
Apimaysin	Not Well Detected	7.29	NA	NA
Caffeic acid	14.18	Not Well Detected	NA	NA
Chrysoeriol	4.57	5.86	1.28	<0.0001
Coniferyl aldehyde	0.17	2.36	13.88	<0.0001
Dihydrokaempferol	0.33	Not Well Detected	NA	NA
Dihydroquercetin	0.24	1.19	4.96	<0.0001
Eriodictyol	0.65	2.90	4.46	<0.0001
Eriodictyol-7-O-glucoside	Not Well Detected	2.46	NA	NA
Ferulic acid	Not Well Detected	7.02	NA	NA
Isoorientin	Not Well Detected	3.90	NA	NA
Kaempferol	Not Well Detected	4.29	NA	NA
Luteolin	0.26	3.05	11.73	<0.0001
Luteolin-7-O-glucoside	Not Well Detected	Not Well Detected	NA	NA
Maysin	Not Well Detected	10.21	NA	NA
Naringenin	0.27	3.44	12.74	<0.0001
Naringin	Not Well Detected	Not Well Detected	NA	NA
Naringenin-7-O-glucoside	Not Well Detected	3.81	NA	NA
p-Coumaric acid	209.77	9.37	0.04	<0.0001
Phenylalanine	35.85	5.68	0.16	<0.0001
Quercetin	Not Well Detected	3.19	NA	NA
Quinic acid	9.91	8.47	0.85	<0.0001
Rhamnosylisoorientin	1.22	6.29	5.16	<0.0001
Shikimic acid	Not Well Detected	4.29	NA	NA
Sinapic acid	0.42	2.22	5.29	<0.0001
Syringic acid	32.56	Not Well Detected	NA	NA
Syringol	Not Well Detected	Not Well Detected	NA	NA
Tricin	14.10	11.43	0.81	<0.0001
Vanillic acid	169.51	8.38	0.05	<0.0001
Vanillin	442.78	8.65	0.02	<0.0001
Vitexin	0.91	5.94	6.53	<0.0001

Although these results provide useful information for the samples and lines studied, this may not show definitive results for all maize kernel and seedling varieties. Although the samples were prepared and run with the same procedures, the post analytical chemistry steps were done differently, with the data cleaning and outlier removal. This may cause discrepancies with some of the results. Additional analysis of the progression of phenolic compound accumulation

between the maize kernel and seedling tissue will provide greater understanding on how the expression of specific genes changes over the maize lifecycle, and the purposes of specific compounds in the maize plants at each stage of growth.

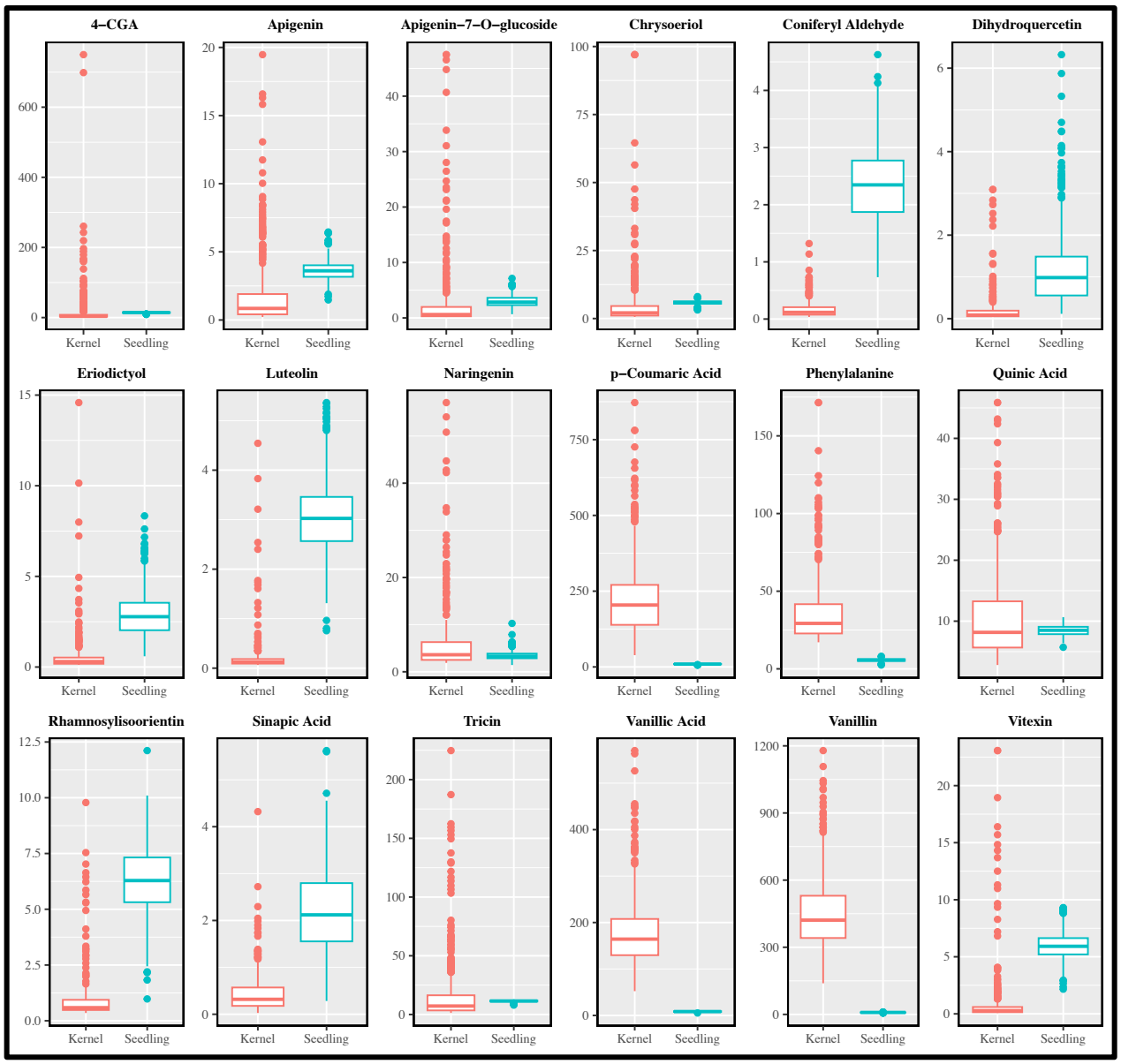


Figure 3.7: Boxplots of the phenolic compounds detected in common between maize kernel and maize seedling tissue samples. All of the y-axes are in the units of AUA, and all x-axes have the boxplot for kernel samples on the left (red) and seedling samples on the right (blue). Both sets of samples went through the same preparation but different cleaning methods. Figure made with the R package ggplot2, (Wickham, 2016).

CONCLUSION

Overall, maize kernel tissue from the Wisconsin Diversity Panel shows a wide variety of phenolic compound accumulation. These results vary based on variety and subpopulation genetic group. Phenolic compound research is essential for understanding the nutritional benefits from consuming plants and the benefits the plants receive due to their production. Future research on deciphering the phenotypic differences corresponding to phenolic compound accumulation can be key to understanding other modes of disease resistance, environmental stressor resistance, and increased nutritional elements. Some of the next steps for this research include exploring the candidate genes found from the GWAS, analyzing the accumulation patterns on mature maize leaves, and comparing field phenotypes with the phenolic compound profiles to determine any connections.

REFERENCES

Ballard, C. R. & Junior, M. R. M. (2019). Chapter 10 - Health Benefits of Flavonoids. In M. R. S. Campos (Ed.), *Bioactive Compounds* (pp. 185–201). essay, Woodhead Publishing. https://doi.org/10.1016/B978-0-12-814774-0.00010-4

Barrière, Y., Alber, D., Dolstra, O., Lapierre, C., Motto, M., Ordás Pérez, A., Van Waes, J., Vlasminkel, L., Welcker, C., & Monod, J. P. (2006). Past and prospects of forage maize breeding in Europe. II. History, germplasm evolution and correlative agronomic changes. *Maydica, 51,* 435-449. http://hdl.handle.net/10261/42855

Bashandy, T., Taconnat, L., Renou, J. P., Meyer, Y., & Reichheld, J. P. (2009). Accumulation of flavonoids in an NTRA ntrb mutant leads to tolerance to UV-C. *Molecular Plant*, *2*(2), 249–258. https://doi.org/10.1093/mp/ssn065

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633-2635.

Bredsdorff, L., Nielsen, I. L., Rasmussen, S. E., Cornett, C., Barron, D., Bouisset, F., Offord, E., & Williamson, G. (2010). Absorption, conjugation and excretion of the flavanones, naringenin and hesperetin from α-rhamnosidase-treated orange juice in human subjects. *British Journal of Nutrition*, *103*(11), 1602–1609. https://doi.org/10.1017/s0007114509993679

Beckett, Travis J. (2016). *Analysis of Genetic Loci Associated with Agronomic Performance in Previously Plant-Variety-Protected Elite Commercial Maize Germplasm*. Open Access Theses. 1182. https://docs.lib.purdue.edu/open_access_theses/1182

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast Two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, *108*(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Chen, S., Wang, X., Cheng, Y., Gao, H., & Chen, X. (2023). A review of classification, biosynthesis, biological activities and potential applications of flavonoids. *Molecules*, *28*(13), 4982. https://doi.org/10.3390/molecules28134982

De La Fuente, G. (2015). *Improvements to the maize (Zea mays L.) in vivo maternal doubled haploid system,* Iowa State University. Graduate Theses and Dissertations. 14767. Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. https://lib.dr.iastate.edu/etd/14767

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package

rrBLUP. Plant Genome 4:250-255.

Erlund, I. (2004). Review of the flavonoids quercetin, hespertin, and naringenin. Dietary sources, bioactivities, bioavailability, and epidemiology. *Nutritional Resources, 24*, 851-874. https://doi.org/10.1016/j.nutres.2004.07.005

Gomez-Cano, L., Rodriguez, J., de Leon, N., and Grotewold, E. (2023). Major determinants of maize natural variation in phenolic compound accumulation. Manuscript in preparation.

Grzybowski, M.W., Mural, R.V., Xu, G., Turkus, J., Yang, J. & Schnable, J.C. (2023) A common resequencing-based genetic marker dataset for global maize diversity. Plant Journal, 113, 1109– 1121.

Gu, Z. (2022). Complex heatmap visualization. *iMeta*, *1*(3). https://doi.org/10.1002/imt2.43

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Hallauer, Arnel R., "Corn Breeding" (2009). Iowa State Research Farm Progress Reports. 475. http://lib.dr.iastate.edu/farms_reports/475

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., Della Coletta, R., Tittes, S., Hudson, A. I., Marand, A. P., Wei, S., Lu, Z., Wang, B., Tello-Ruiz, M. K., Piri, R. D., … Dawe, R. K. (2021). De Novo Assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, 373(6555), 655–662. https://doi.org/10.1126/science.abg5289

Kassambara, A. (2022). _ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'_. R package version 0.1.4. https://CRAN.R-project.org/package=ggcorrplot

Lattanzio, V., Cardinali, A., Linsalata, V. (2012). Plant phenolics: a biochemical and physiological perspective. In: Cheynier, V., Sarni-Manchado, P., Quideau, S. (Eds.), *Recent Advances in Polyphenol Research*. John Wiley & Sons, Ltd, pp. 1–39. https://doi.org/10.1002/9781118299753.ch1

Lenth, R. (2023). _emmeans: Estimated Marginal Means, aka Least-Squares Means_. R package version 1.8.4-1, <https://CRAN.R-project.org/package=emmeans>.

Liu, K., Goodman, M., Muse, S., Smith, J.S., Buckler, E., and Doebley, J. (2003). Genetic Structure and Diversity Among Maize Inbred Lines as Inferred From DNA Microsatellites. *Genetics, 165*(4), 2117-2128. https://doi.org/10.1093/genetics/165.4.2117

Liu, W., Feng, Y., Yu, S., Fan, Z., Li, X., Li, J., & Yin, H. (2021). The Flavonoid Biosynthesis Network in Plants. *International Journal of Molecular Sciences*, *22*(23), 12824.

https://doi.org/10.3390/ijms222312824

Liu, Z., Liu, Y., Pu, Z., Wang, J., Zheng, Y., Li, Y, & Wei, Y. (2013). Regulation, evolution, and functionality of flavonoids in cereal crops. *Biotechnology Letters 35*, 1765–1780. https://doi.org/10.1007/s10529-013-1277-4

Lu, H., & Bernardo, R. (2001). Molecular marker diversity among current and historical maize inbreds. *Theoretical and Applied Genetics*, *103*(4), 613–617. https://doi.org/10.1007/pl00002917

Mazaheri, M., Heckwolf, M., Vaillancourt, B., Gage, J.L., Burdo, B., Heckwolf, S., Barry, K., Lipzen, A., Ribeiro, C.B., Kono, T.J.Y., Kaeppler, H.F., Spalding, E.P., Hirsch, C.N., Buell, C.R., de Leon, N., & Kaeppler, S.M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol 19, 45 https://doi.org/10.1186/s12870-019-1653-x

Mikel, M.A. & Dudley, J.W. (2006). Evolution of North American Dent Corn from Public to Proprietary Germplasm. *Crop Sci.*, 46: 1193-1205. https://doi.org/10.2135/cropsci2005.10-0371

Mural, R. V., Sun, G., Grzybowski, M., Tross, M. C., Jin, H., Smith, C., Newton, L., Andorf, C. M., Woodhouse, M. R., Thompson, A. M., Sigmon, B., & Schnable, J. C. (2022). Association mapping across a multitude of traits collected in diverse environments in maize. *GigaScience*, 11. https://doi.org/10.1093/gigascience/giac080

Ngapo, T. M., Bilodeau, P., Arcand, Y., Charles, M. T., Diederichsen, A., Germain, I., Liu, Q., MacKinnon, S., Messiga, A. J., Mondor, M., Villeneuve, S., Ziadi, N., & Gariépy, S. (2021). Historical indigenous food preparation using produce of the three sisters intercropping system. *Foods*, *10*(3), 524. https://doi.org/10.3390/foods10030524

Panche, A. N., Diwan, A. D., & Chandra, S. R. (2016). Flavonoids: an overview. *Journal of Nutritional Science, 5*(47). https://doi.org/10.1017/jns.2016.41

Pandey, S., & Gardner, C. O. (1992). Recurrent selection for population, variety, and hybrid improvement in tropical maize. *Advances in Agronomy*, 1–87. https://doi.org/10.1016/s0065-2113(08)60935-9

Parker, A. C. (1910). *Iroquois uses of maize and other food plants*. Univ. of the State of New York.

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

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., & Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences*, *98*(20), 11479–11484. https://doi.org/10.1073/pnas.201394398

Rodriguez, J., Gomez-Cano, L., Grotewold, E., & de Leon, N. (2022). Normalizing and correcting variable and complex LC–ms metabolomic data with the R package pseudodrift. *Metabolites*, *12*(5), 435. https://doi.org/10.3390/metabo12050435

Rouf Shah, T., Prasad, K., & Kumar, P. (2016). Maize – A potential source of human nutrition and Health: A Review. *Cogent Food & Agriculture*, *2*(1). https://doi.org/10.1080/23311932.2016.1166995

Šamec, D., Karalija, E., Šola, I., Bok, V. V., & Salopek-Sondi, B. (2021). The role of polyphenols in abiotic stress response: The influence of molecular structure. *Plants, 10*(1)*,* 118. https://doi.org/10.3390/plants10010118

Soto-Vaca, A., Losso, J. N., Xu, Z., & Finley, J. W. (2021). Review: Evolution of phenolic compounds from color and flavor problems to health benefits. *Journal of Agriculture and Food Chemistry, 60*(27), 6658-6677. https://doi.org/10.1021/jf300861c

Stopsack, K. H., Tyekucheva, S., Wang, M., Gerke, T. A., Vaselkiv, J. B., Penney, K. L., Kantoff, P. W., Finn, S. P., Fiorentino, M., Loda, M., Lotan, T. L., Parmigiani, G., & Mucci, L. A. (2021). Extent, impact, and mitigation of batch effects in tumor biomarker studies using tissue microarrays. *eLife*, *10*. https://doi.org/10.7554/elife.71265

Swanson, B.G. (2003). Tannins and Polyphenols. *Encyclopedia of Food Sciences and Nutrition (Second Edition) Academic Press.* 5729-5733. https://doi.org/10.1016/B0-12-227055-X/01178-0

Tracy, W. F. (2001). Sweet Corn. In A. R. Hallauer (Ed.), *Specialty corns* (Second Edition, pp. 155–198). essay, CRC Press.

Wang J., Zhang Z., GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction, Genomics, Proteomics & Bioinformatics (2021). https://doi.org/10.1016/j.gpb.2021.08.005.

Waters Corporation. (2021). *MassLynx Mass Spectrometry Software*. Waters. https://www.waters.com/waters/en_US/MassLynx-Mass-Spectrometry-Software

Waters Corporation. (2021). *TargetLynx*. Waters. https://www.waters.com/waters/en_US/TargetLynx-/nav.htm?locale=en_US&cid=513791

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Isbn: 978-3-319-24277-4, https://ggplot2.tidyverse.org

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, Cell Biology, and biotechnology. *Plant Physiology*, *126*(2), 485–493. https://doi.org/10.1104/pp.126.2.485

Xiao, Z. P., Peng, Z. Y., Peng, M. J., Yan, W.-B., Ouyang, Y.-Z., & Zhu, H.-L. (2011). Flavonoids health benefits and their molecular mechanism. *Mini-Reviews in Medicinal Chemistry*, *11*(2), 169–177. https://doi.org/10.2174/138955711794519546

Xu, D. P., Li, Y., Meng, X., Zhou, T., Zhou, Y., Zheng, J., Zhang, J. J., & Li, H. B. (2017). Natural Antioxidants in Foods and Medicinal Plants: Extraction, Assessment and Resources. *International Journal of Molecular Sciences. 18*(1), 96. https://doi.org/10.3390/ijms18010096

Zhang, Z., Liang, Z., Yin, L., Li, Q. X., & Wu, Z. (2018). Distribution of Four Bioactive Flavonoids in Maize Tissues of Five Varieties and Correlation with Expression of the Biosynthetic Genes. *Journal of Agriculture and Food Chemistry, 66*(40), 10431-10437. https://doi.org/10.1021/acs.jafc.8b03865

Ziegler, K. E. (2003). *Popcorn production and marketing*. Division of Agricultural Sciences and Natural Resources, Oklahoma State University.

CHAPTER 4: UTILIZING FOURIER TRANSFORM MID-INFRARED SPECTROSCOPY TO MODEL PHENOLIC COMPOUND ACCUMULATION IN DIVERSE MAIZE KERNEL TISSUE

ABSTRACT

Phenotyping is essential for the advancement of plant breeding and results in numerous successful crop improvement efforts. Many widely used methods for phenotyping are time consuming and costly, making the resulting data inaccessible for those without the funding or substantial equipment needed to produce the data. Phenolic compounds accumulation in plants is an important phenotype due to their role in plant defense and their high nutritive properties. Traditionally, to detect and quantify phenolic compounds in plant tissues, extensive wet lab and analytical chemistry needs to be done. In this study, we explore how FT-MIR spectroscopy can be used to model these phenotypes in maize kernel tissue from a subset of lines in the Wisconsin Diversity Panel. Using the ground kernel tissue samples with FT-MIR spectroscopy, coupled with the statistical modeling methods, random forest and partial least squares regression, allow the investigation into the feasibility of this phenotyping method. Exploring a variety of spectral preprocessing methods within each model type, that are commonly used in soil sciences, allowed the exploration of varying levels of prediction accuracy of the models. Implementing different techniques will allow us to continue to optimize the accuracy of each model for the best phenolic compound accumulation prediction. Eventually, this research will provide an insightful and applicable tool for breeders and researchers to model phenotypes that are originally difficult to quantify, with spectral response data.

INTRODUCTION

Plant breeding has been an essential part of history, resulting in numerous advancements in order to continue to support the growing planet. In order to continue to increase the quantity and quality of the crops produced, new techniques are needed to aid and advance plant breeding further. One of the bottlenecks of plant breeding, is the time-consuming nature of phenotyping. Plant phenotyping has been a means of variety selection for hundreds of years, and is the original method used for crop domestication (Kumar et al., 2015). Phenotyping is essential for the continued progression of breeding pipelines and the improvement of crops.

Phenotyping for phenolic compound accumulation includes costly, labor intensive, time consuming wet lab chemistry that makes this technique difficult to employ on a large scale and inaccessible for many to use. Previous methods of quantifying and detecting phenolic compound accumulation in plant tissues is done by chemical extraction and analytical chemistry through the use of liquid chromatography – mass spectrometry (LC-MS). LC-MS is an analytical chemistry technique that combines the physical separation granted achieved through liquid chromatography with the mass analysis and quantification properties of mass spectrometry (Cocuron et al., 2019). The laborious wet lab and analytical chemistry involved in phenolic compound extraction and LC-MS reduces the accessibility of the data collection and prevents the widespread use. Moving away from traditional extraction and quantification methods and towards a phenomic type approach could lead to a reliable, efficient and affordable method to decipher mass amounts of data without the constraints of the previous methods (Kumar et al., 2015; Houle et al., 2010).

Spectroscopy is the study of the reflectance and absorption of light by matter and it is a technique that involves splitting light, or electromagnetic radiation, into a spectrum of wavelengths, to gain more knowledge on the properties of the matter (Dutta, 2017). When used

to gain information on a sample, some of the electromagnetic radiation is transmitted, reflected, and absorbed. The resulting spectral response is unique to the components of the sample creating a fingerprint of the contents (Dutta, 2017). Spectroscopy is known for its high throughput nature and gathering large data that can be used for a wide range of applications (Kalendar et al., 2022).

Fourier transform – infrared (FT-IR) spectroscopy is a technique that utilizes infrared light to scan and observe the chemical properties and makeup of samples (Berna, 2017). This technique captures information on how IR light changes the dipole moments in molecules and the reactions to specific vibrational energy (JASCO inc., 2023). FT-IR spectroscopy has passed historical phenotyping techniques because it can test all of the wavelengths of IR light at the same time leading to a much faster process (Bruker, 2023). One of the most popular applications of FT-IR spectroscopy is in the use of soil sciences. It is a technique that is used readily for identifying and characterizing complex organic macromolecules within the soil (Stevenson, 1982). An additional important application of FT-IR spectroscopy is the nutrient analysis of plant tissues and their products. An important study found success in confirming the presence of phenols, alkanes, amines, carboxylic acids, nitrile, aromatics, and alcohols which helped uncover information about the nutritional elements in the plant tissues including proteins, vitamins, carbohydrates and amino acids (Bachhar et al., 2023). Another research study, done by Mierzwa-Herszetek et al. (2019), was able to successfully detect the total phenolic content of biochar through the analysis of exogenous organic matter from plant biomass. These studies showed promising results that propose the ability to reduce the costly and time-consuming wet lab chemistry with FT-IR spectroscopy. Applying spectral preprocessing and analysis methods that are primarily used in soil sciences could create potential for rapid phenotyping of chemical

components of plant tissues, removing the need for time consuming and costly historical methods.

Coupling the spectral response from FT-IR spectroscopy with traditional phenotyping data, provides the potential for the use of modeling to predict phenotypes accurately. Developing and selecting the correct model type depends on the structure of the data, number of samples, potential of overfitting, and assessment of performance (Núñez et al., 2011). Both random forest and partial least squares are regression modeling techniques commonly used in research for effect estimation and prediction, especially with large data like spectral. The random forest modeling method is categorized as a machine learning algorithm, that uses decision trees that lead to either categorical or continuous outcomes (Breiman, 2001; Schonlau & Zou, 2020). Partial least squares regression analysis is a statistical approach that permits researchers to compare multiple response and explanatory variables. This is completed by combining features from principal component analysis and multiple linear regression (Abdi, 2003). Although random forest and partial least squares are both modeling techniques, they have different algorithms used for decision making and prediction which could create different model quality and analysis.

Although many techniques have been explored for the analysis of soils and plant properties as a whole, there is less research done on the modeling of individual phenolic compounds through spectral response. In this study, we aim to: (1) explore the connection between FT-IR spectral response and phenolic compound accumulation in maize tissues, (2) employ numerous preprocessing techniques to the spectral response to determine if there is a superior method that increases the quality of the spectra before analysis, (3) utilize both random forest and partial least squares regression modeling techniques to predict phenolic compound

accumulation phenotypes from spectral response data, and (4) to evaluate which modeling technique is most accurate for this type of data.

MATERIALS AND METHODS

Plant materials and experimental design

Maize kernels were sourced from GRIN Global and are all a part of the Wisconsin Diversity panel. These kernels were harvested upon maturity and had the correct kernel moisture to indicate this. 100 different maize varieties were selected from the panel as a way to represent the range of diversity and to provide diverse phenotypes. Three different kernels were selected from each variety and were used as a means of biological replicates. A list of the varieties and sample names used in this experiment can be seen in the supplemental files. Individual varieties were placed into paper labeled envelopes to reduce error and maintain organization.

Tissue preparation

The maize kernels, in their paper envelopes, were lyophilized for 5-7 days to complete the freeze-drying process and to ensure lower moisture levels. Excess moisture in the kernels may affect the dry weights and normalization procedures which will affect the final phenotype readings of the phenolic compounds. After the kernels were adequately freeze dried, they were placed into reinforced 2 mL bead mill tubes with three steel 5 mm beads. The kernels were then run through the Bead Mill tissue homogenizer. The tissue homogenizer was set at a speed of 5.00, a grind time of 8 seconds, a dwelling time of 9 minutes and 59 seconds, and was run for 6 cycles. The grind time is set for a long length of time to prevent the tubes and tissue from heating up. After the 6 cycles are finished, the kernels are inspected for a fine powder consistency. If the

kernel tissue is not ground into a fine powder, or there are visible chunks of tissue, the samples are sent back through the tissue homogenizer for three additional cycles. This is repeated until there are no viable chunks of tissue and the samples are ground into a fine powder. 29.0 – 31.0 mg of kernel tissue was allocated into 1.5 mL microcentrifuge tubes and weight measurements were recorded to the nearest tenth of a milligram. The remnant ground tissue powder was kept in the reinforced tubes until ready for use.

Methanol extraction of phenolic compounds

A solution of 80% MeOH and 0.1% formic acid was prepared and 500 µL was added to each sample. The sample tubes were then vortexed and placed in a 4° C refrigerator, undisturbed for at least 12 hours, or overnight. After the period in the refrigerator, the samples were then vortexed again and centrifuged at 13000 rpm for 3 minutes. Then, 400 µL of the supernatant was collected and reallocated into a clean 1.5 mL microcentrifuge tube. The samples were then diluted with autoclaved pure water to obtain a 50% MeOH solution concentration. The samples were then vortexed and stored in the 4° C refrigerator until use.

Analytical chemistry to detect and quantify phenolic compound composition

The instrument used for the LC-MS analytical chemistry analysis was a Waters ACQUITY TQD Tandem Quadrupole UPLC/MS/MS (Waters Corporation, Milford, MA, USA). The maize kernel extract samples were run through the instrument alongside a standard calibration curve, pool quality control samples, blanks, and an internal standard. 90 µL of each sample was pipetted into 96 well plates. An internal standard of 10 µL of 500 nM 8-prenylnaringenin was added into each well of the plate to bring the internal standard

concentration to 50 nM. The 96 well plates were run after a set of amber vials comprised of the calibration curve. This was done before each of the 96 well plates for higher quality samples.

The phenolic compound standards used to make the calibration curve were sourced from Sigma-Aldrich (Burlington, MA, USA) Cayman Chemical (Ann Arbor, MI, USA), Indofine chemical (Hillsborough, NJ, USA), and ChromaDex (Irvine, CA, USA) except apimaysin, maysin, and rhamnosylisoorientin, which were provided by Michael McMullen (USDA-ARS) and Maurice Snook (Iowa State University) facilitated through Dr, Erich Grotewold (MSU), information on the standards can be found in Rodriguez et al. (2022). The calibration curve was made up of 34 phenolic compound standards in a series of serial dilutions including 3.9 nM, 15.625 nM, 31.25 nM, 62.5 nM, 125 nM, 250 nM, 500 nM, 750 nM, and 1000 nM. The calibration curve of standards was pipetted into amber vials and run before every plate. The TQD Tandem Quadrupole UPLC/MS/MS was run on a targeted 10-minute targeted multiple reaction monitoring method setting to detect and quantify the compounds against the standard calibration curve. The targeted multiple reactions monitoring methods file was modified from Rodriguez et al. (2022).

Data normalization and preparation

Relative values derived from the integrated peak area in the produced chromatogram from the LC-MS data acquisition were used instead of absolute area based on the range of concentrations in the standard calibration curve. To combat noise and machine error, phenolic compounds in each sample were only categorized as detectable if the area of the chromatogram was greater than three times the mean area of the blank samples. Any samples that did not meet

the threshold, were considered to not meet the limits of detection and were given a value of NA for that phenolic compound.

Normalization of the samples was conducted using a calculation of arbitrary units of area (AUA). AUA was calculated with each phenolic compound in each sample separately, by dividing the area of the integrated peak by milligram of kernel tissue per mL of methanol solution. Greater detail on the calculation of AUA can be seen in Rodriguez et al. (2022). This calculation accounted for the variation in weight of the ground kernel tissue during the weighing sample preparation after the tissue grinding. Outliers were removed using linear modeling that takes into account maize variety and Cook's distance. The model used was made through the 'lm' function in R and cook's distance was calculated using the R function 'cooks.distance'. Data points were removed and replaced with NA if they were considered influential and the cook's distance of the point was greater than 4 divided by the sample size. This was done individually for each phenolic compound.

Microplate preparation and FT-IR spectra acquisition

The remnant ground tissue samples from the same kernels used in the wet lab and analytical chemistry analysis was used to fill 96 well microplates. Each maize variety had 3 biological replicates, 3 different kernels, and 2 different technical replicates for each kernel, to total 6 samples/wells per maize line. To preserve the cleanliness and organization this research needed, glass plate covers were used to cover the wells not being used. A metal spatula, cleaned with isopropyl alcohol, was used to scoop ground kernel tissue from the 2 mL tube into a well. Enough kernel tissue was used to cover the bottom of the well so there was no visible metal showing through. A metal peg tool, cleaned with isopropyl alcohol was then used to flatten down

the well and pack the tissue together tightly. This removes any large height discrepancies from the well. If there were any chunks of kernel tissue, they were removed and only the fine powder was used. A small suction vacuum was then used to remove any excess powder from the microplate, and a glass cover was placed over the well. The tools are cleaned with isopropyl alcohol between each well.

After the plates were adequately prepared and quality checked for level packed wells, they were run through the Bruker Vertex 70v FT-IR spectrometer (Bruker Optics, Billerica, MA USA). This tool was used to obtain the FT-MIR spectral response from the maize kernel samples. The biological and technical replicates for the same variety were given identification names that included which variety, kernel, and replicate the sample well was a part of.

Spectral preprocessing methods

Receiving the spectra from the output of the Bruker Vertex 70v FT-IR spectrometer is in its raw form. The raw form of the spectral data can have machine noise and errors that deplete the quality of the analysis. Using spectral pre-processing methods, leads to improving the quality of the spectra before the analysis and removing any scattering noise (Wadoux et al., 2021). Using the r packages prospectr, signal, and pracma numerous pre-processing methods were created from the raw spectra (Stevens & Ramirez-Lopez, 2022; Signal Developers, 2014; Borchers, 2022). The raw spectra was extracted and evaluated using the Bruker software OPUS (Bruker Optics, Billerica, MA USA). This software was used to extract the spectra from the spectrometer and format the data into a usable file to then be read into R. In figures, tables and supplemental files, the raw spectrum is denoted as spc.

The first of the preprocessing methods was the spectral resampling technique. This technique reduces the dimensions of the spectra so that there are less wavelengths to consider (Wadoux et al., 2021). This process uses an interpolation to resample the spectra to a set of new coordinates and at a different resolution. For this research, the spectrum was trimmed from wavenumbers of 600 cm^{-1} to 4000 cm^{-1} and was resampled for every 2 cm^{-1} wavenumbers to give 1700 different sample points. The spectrum resampling method are abbreviated as spcrs.

Another of the preprocessing techniques includes using a moving window average function from the proscpectr package (Stevens & Ramirez-Lopez, 2022). The moving window average calculates the average of the neighboring wavenumbers within the specified window size from the original spectra. This creates a smoothing effect across the entire spectra and limits the excess noise (Wadoux et al., 2021). The moving widow average technique was used on both the raw and resampled spectra with different window sizes of 5, 10, 11, 15, and 20 cm^{-1}; these techniques are annotated by the starting spectra, the acronym MWA, and the window size (the raw spectra with a window size of 5 is 'spcMWA5', the resampled spectra with a window size of 20 is 'spcrsMWA20'). An additional smoothing preprocessing method is called Savitzky-Golay filtering, annotated as 'SG'. This method uses a polynomial regression, of a user specified order, on a series of spectral values which will determine the smoothed value for each wavelength for a filter length/window size (Wadoux et al., 2021). This method uses the R package signal and the 'filterSg' function outlined in Wadoux et al (2021) with a window size of 11 wavenumbers and a second order polynomial (Signal Developers, 2014).

An alternative method is using the standard normal variate transform, or SNV. SNV corrects for single light scattering through using a z-transformation to center and scale the spectrum (Wadoux et al., 2021). This was done using the prospectr R package with the function

‘standardNormalVariate’ (Stevens & Ramirez-Lopez, 2022). This was done on the resampled spectra, but it can also be done on spectra that has already been filtered or smoothed. Next, a multiplicative scatter correction method was used to account for multiplicative deviations that are dependent from the wavelengths (Wardoux et al., 2021). This method uses an alignment correction to a reference spectrum to the amplification effects are at the same average in every spectrum (Wardoux et al., 2021). The spectrum is then fit using the least squares method. This was done using the ‘msc’ function from the proscpectr package in R and using the wavelength means as the reference spectrum (Stevens & Ramirez-Lopez, 2022). This method is labeled as MscC. An alternative method to standard normal variate and multiplicative scatter correction is using a detrending method, this can be used in addition to the other methods or by itself. Using the proscpectr package in R, the ‘detrend’ function is used which involves using an standard normal variate transformation with a second order polynomial to return the residuals (Stevens & Ramirez-Lopez, 2022; Wardoux et al., 2021).

Spectral standardization is another preprocessing method used to transform the spectral response values to zero mean and unit variance (Wardoux et al., 2021). Standardization is done by subtracting each spectral wavenumber by the mean of all the spectra values for the wavenumber and dividing by the standard deviation (Wardoux et al., 2021). Spectral centering or normalization is used to transform the spectral values in each wavenumber or wavelength to zero mean. Centering is done by subtracting the spectral wavenumber by the mean of all spectra values (Wardoux et al., 2021). Both of these techniques can be done with the base package in R using the ‘scale’ function. Spectral standardization is abbreviated as Sdt and spectral centering is labeled as Norm.

Another method for preprocessing the spectra is the use of derivatives. Transforming the spectra into first or second order derivatives results in highlighting the contents absorption features (Wardoux et al., 2021). The first order derivative is used mostly to detrend the spectrum and the second order derivative is used to both detrend the spectrum and removing any linear trends (Wardoux et al., 2021). This was done by building on the Savitsky-Golay 'filterSg' function by specifying the order of the derivative. This method was done on both the resampled and the raw spectra and the acronym indicates the spectra used with the order of the derivative taken (Deriv1 or Deriv2).

Lastly, the continuum removal (CR), or convex hull, technique is used to spotlight absorption or reflectance features of sample components, especially minerals (Stevens & Ramirez-Lopez, 2022). The continuum removal technique assigns values between 0 and 1 to the specified regions and can be done on both reflectance and absorption data. This was done on the resampled data and is given the abbreviation cr. A table of all the preprocessing methods, their labels in figures and tables, and a brief description can be found in table 4.1.

Table 4.1: List of the preprocessing methods used on the kernel spectra. This table also includes the labels used in other figures and supplementary files as well as a brief description of the preprocessing method. Information on the preprocessing methods was derived from Wardoux et al. (2021).

Preprocessing Method	Label	Description
Raw Spectra	scp	- Unadjusted spectral response with no preprocessing methods and no trimming
Spectral Resampling	scprs	- Trimmed to only include the MIR range and resampled at every 2 wavenumbers (cm^{-1})
Savitzky-Golay Filtering	SG	- Used to smooth a spectrum - Fits a polynomial regression on a series of spectral values to derive a smoothed value for each wavenumber
Standard Normal Variate	SNV	- Corrects for single light scattering through centering and scaling - Done to normalize each spectrum to zero mean and unit variance by subtracting the spectrum mean and dividing by the standard deviation
Multiplicative Scatter Correction	spcrsMscC	- Used to compensate for multiplicative deviations that are depended from wavenumber - Done by the alignment to a reference spectrum so the baseline and amplifications are at the same level
Spectral Detrending	DT1 or DT2	- Removes the mean value or linear trend from the spectra - Two different detrending functions give DT1 and DT2
Spectral Centering	spcrsNorm	- Completed by subtracting the wavenumber value by the mean of all spectral values for this wavenumber
Spectral Standardization	spcrsSdt	- Standardizes through centering and scaling the spectra
First Derivative	Deriv1	- Detrends the spectra - Highlights the regions of absorbance
Second Derivative	Deriv2	- Detrends and removes linear trends from the spectra - Highlights the regions of absorbance
Moving Window Average	MWA	- Each wavenumber value is taken as an average of the surrounding wavenumbers based on a specified window size - Provides a smoothing effect that removes noise
Continuum Removal	CR	- Fits a convex hull to each spectrum and computes the deviations from the hull - Gives a value of 0 to all parts of the absorption spectrum that lie on the convex hull and values between 0 and 1 to regions inside the absorption bands

Modeling techniques

After the spectra is preprocessed with numerous different methods, modeling needs to be done to determine if the spectral response can be used to predict phenolic compound accumulation in the maize kernel tissue. This was done using two different types of models. The first is a random forest model and the second is a partial least squares regression type model.

Random Forest

To create the model, the package randomForest in R was used (Liaw & Wiener, 2002). This was done by first creating a variable with the phenolic compound names and a variable with the preprocessing technique names, both in lists. Next, an output file is created with a column with the phenolic compounds and the preprocessing techniques as the column names, one column for each technique. Then, a nested loop was created that loops through the phenolic compounds within each of the preprocessing techniques. Within the nested loops, the data is subset to include the maize variety names, the single column of specified phenolic compound data, and the specified spectra. Then an 80/20 split for the data was used for training and testing with a 10-fold cross validation method. Within this model, the predictor data, or the spectra is subset so only the top predictors, or most important wavenumbers are used based on the lowest root mean squared error (RMSE). The size of the model based on the number of predictors is limited to only account for 1 through 20. Using the important predictors and the model that created the lowest RMSE, it is used on the test data, and the R2 and correlation is computed as a measure of model quality. Computation for modeling was supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University.

Partial least squares regression

A partial least squares regression model was created using the R packages pls and plsVarSel (Liland et al., 2023; Mehmood et al., 2012). To prepare for this model, two variables were created that held the names of the phenolic compounds and the names of the spectra for each preprocessing technique. Then, empty data frames were created that would eventually be filled with the output statistics as a way to measure the success and accuracy of the model, the statistics include RMSE, MSE, R^2, correlation between observed and predicted data, and the number of components that optimized the model. Next, similarly to the random forest model, a nested loop was created to go through each of the phenolic compounds for each of the preprocessing techniques; each combination of phenolic compound and preprocessing technique was done with a separate model. Within this loop, the data was separated into training and testing data with a 80/20 split. The model was then created using the 'plsr' function in the pls R package with a cross validation method and up to 20 components accounted for (Liland et al., 2023). Then, the RMSEP function was used on the model to list the RMSE for each number of components accounted for within the model. Then the number of components that minimizes the RMSE is selected for and used to test the model and predict the response data. Then the loop outputs the summary statistics and populates the once empty data frames for future evaluation.

RESULTS

Comparing R^2, correlation coefficients, and RMSE for Random Forest and Partial least squares regression modeling

On average, using a random forest model with any of the preprocessing techniques results in an R^2 value of 0.247 greater than the same preprocessing technique with a partial least squares regression model. The random forest models also result in a higher correlation value of 0.241 and a lower RMSE by -2.201. The only occurrences where PLSR resulted in a higher model performance than random forest is with the compounds apigenin, quinic acid, and rhamnosylisoorientin with the preprocessing techniques of the second derivative of both the raw and resampled spectra. In almost all other scenarios, random forest performed better or comparable to PLSR. Overall, the random forest models had higher R^2, higher correlation coefficients, and lower RMSE when compared to PLSR, leading to a higher prediction accuracy and a better performing model.

Preprocessing methods

The raw spectral responses were extracted from the samples using the Bruker Vertex 70v FT-IR spectrometer and Bruker's OPUS software (Bruker Optics, Billerica, MA USA). The spectral response represents the absorbance of the electromagnetic radiation at each wavenumber (cm^{-1}). The resampled spectra differ greatly from the raw spectra, it has a large portion of the noise removed and it is trimmed down to only include the MIR range. This was done through resampling at every 2 wavenumbers and only including the range of 4000 – 600 cm^{-1}. A comparison of the raw and resampled spectra can be seen in figure 4.1.

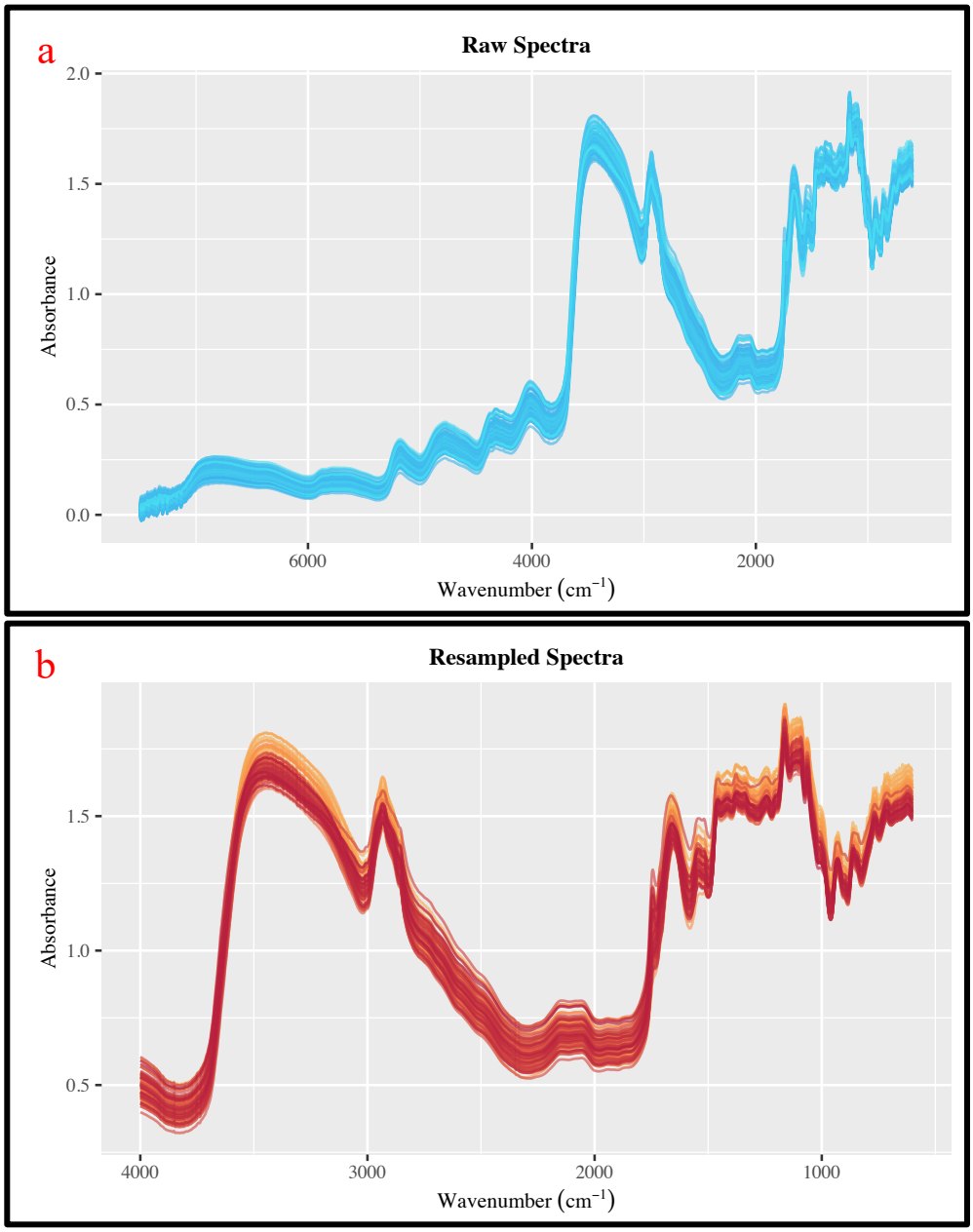


Figure 4.1: Comparison of the spectral responses of all maize samples plotted by absorbance against wavenumber (cm^{-1}). The top plot (a) shows the raw spectral response, while the bottom plot (b) shows the spectral response after being trimmed between wavenumbers 4000 and 600 and resampled to every other whole wavenumber. The figure was created using the R package ggplot2 (Wickham, 2016).

The preprocessing techniques were completed on both the raw and resampled spectra when it was applicable to give a wide variety of techniques to test the models on. Some of the spectral responses only show subtle differences, but these go to impact the model statistics in

significant ways. A plot overviewing the model success through comparing the R^2 values for each of the preprocessing techniques can be seen in figure 4.2**.** Due to the superior results of the random forest modeling when compared to PLSR, the individual model results will only be specified for the random forest techniques. A color-coded table of the R^2 values for the random forest preprocessing techniques can be seen in table 4.2 with a color scale indicating the minimum and maximum model types for each phenolic compound.

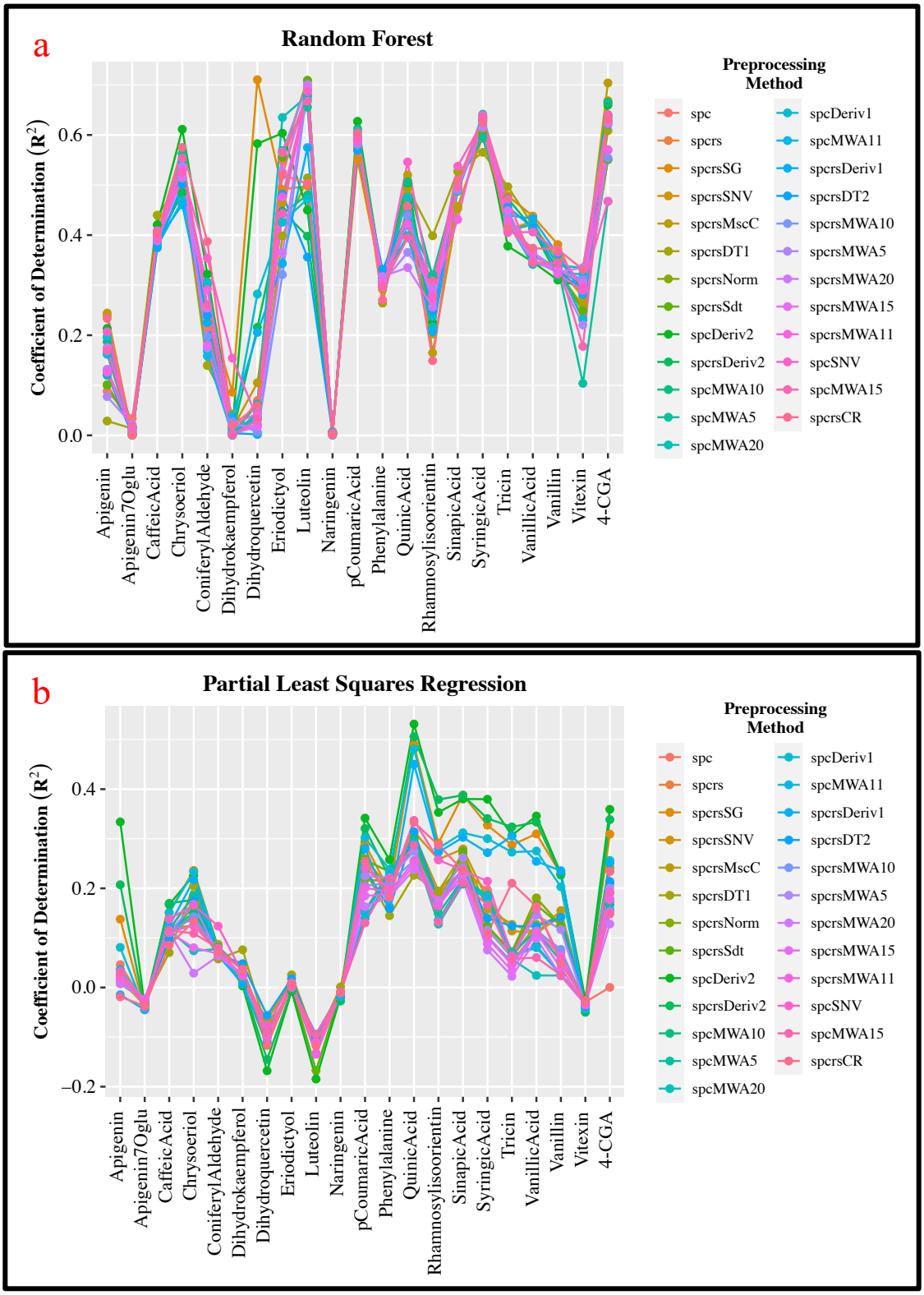


Figure 4.2: R^2 values for all variations of (a) random forest and (b) partial least squares regression models. Color indicates the different preprocessing methods. Colored lines are used to connect the points representing the same preprocessing method for different phenolic compounds. Figure made using the R package ggplot2 (Wickham et al., 2016).

The raw spectra only performed the best with apigenin-7-O-glucoside, but this compound had no models that performed well due to the limited sample size and low values. Using the standard normal variate technique on the raw spectra performed best for dihydrokaempferol, quinic acid, and sinapic acid. The second derivative of the raw spectra performed the best for chrysoeriol. The moving window average with a window of 20 on the raw spectra resulted in the highest R^2 for eriodictyol and syringic acid.

The resampled spectra showed the highest result for luteolin. Savitzky-Golay smoothing resulted in the highest R^2 in dihydrokaempferol, naringenin, vanillic acid, and vanillin. Multiplicative scatter correction worked the best for 4-CGA and apigenin. Both methods of spectral detrending of the resampled spectra showed the highest results in caffeic acid, p-coumaric acid, rhamnosylisoorientin tricin, and phenylalanine. Spectral centering and spectral normalization of the resampled spectra performed the exact same for every compound and showed the highest R^2 in luteolin, in addition to the resampled spectra. The moving window average with a window size of 20 on the resampled spectra was the best technique for vitexin and the continuum removal method was best in coniferyl aldehyde. 14 of the phenolic compounds performed best on models that use the resampled spectra, while 7 of the compounds performed best with preprocessing techniques on the raw spectra.

Table 4.2: R^2 values for the random forest model with all the different preprocessing techniques for each phenolic compound. Color scale is done column wise with the red value in each column representing the highest R^2 value, the blue value representing the minimum R^2 value, and white representing the 50% midpoint value of the column.

Starting Spectra	Preprocessing Method	4-CGA	Apigenin	Apigenin-7-O-glucoside	Caffeic Acid	Chrysoeriol	Coniferyl Aldehyde	Dihydrokaempferol	Dihydroquercetin	Eriodictyol	Luteolin	Naringenin	pCoumaric Acid	Phenylalanine	Quinic Acid	Rhamnosylisoorientin	Sinapic Acid	Syringic Acid	Tricin	Vanillic Acid	Vanillin	Vitexin
Raw	spc	0.633	0.088	0.034	0.392	0.556	0.216	0.001	0.069	0.549	0.683	0.004	0.593	0.306	0.444	0.278	0.509	0.593	0.428	0.343	0.335	0.231
	SNV	0.468	0.207	0.003	0.392	0.525	0.354	0.154	0.033	0.443	0.688	0.002	0.592	0.295	0.546	0.275	0.537	0.626	0.405	0.406	0.366	0.296
	Deriv1	0.635	0.197	0.002	0.378	0.469	0.306	0.010	0.282	0.426	0.472	0.003	0.609	0.308	0.476	0.237	0.491	0.634	0.409	0.422	0.340	0.232
	Deriv2	0.660	0.214	0.003	0.421	0.611	0.322	0.016	0.583	0.604	0.450	0.004	0.627	0.312	0.504	0.226	0.498	0.631	0.378	0.346	0.310	0.296
	MWA5	0.467	0.120	0.016	0.392	0.562	0.194	0.000	0.045	0.569	0.656	0.003	0.587	0.312	0.440	0.321	0.504	0.594	0.439	0.347	0.336	0.104
	MWA10	0.665	0.187	0.017	0.383	0.548	0.254	0.004	0.034	0.443	0.480	0.003	0.591	0.308	0.506	0.316	0.496	0.633	0.444	0.358	0.332	0.299
	MWA11	0.639	0.162	0.014	0.398	0.548	0.159	0.002	0.037	0.477	0.671	0.000	0.595	0.303	0.427	0.261	0.496	0.634	0.427	0.351	0.325	0.322
	MWA15	0.641	0.169	0.018	0.403	0.554	0.260	0.001	0.031	0.565	0.667	0.004	0.601	0.298	0.395	0.308	0.496	0.637	0.472	0.345	0.344	0.178
	MWA20	0.621	0.169	0.019	0.382	0.545	0.169	0.000	0.063	0.635	0.678	0.005	0.596	0.311	0.399	0.217	0.499	0.641	0.425	0.341	0.340	0.334
Resampled	spcrs	0.551	0.101	0.014	0.381	0.531	0.253	0.006	0.020	0.555	0.709	0.004	0.591	0.304	0.424	0.247	0.505	0.602	0.429	0.361	0.333	0.248
	SG	0.627	0.162	0.000	0.388	0.487	0.259	0.086	0.710	0.491	0.498	0.007	0.551	0.301	0.482	0.202	0.497	0.626	0.478	0.438	0.381	0.271
	SNV	0.668	0.238	0.000	0.391	0.515	0.294	0.005	0.061	0.520	0.692	0.007	0.574	0.295	0.492	0.237	0.457	0.597	0.411	0.427	0.338	0.276
	MscC	0.704	0.244	0.000	0.380	0.563	0.227	0.013	0.105	0.465	0.677	0.007	0.585	0.291	0.520	0.165	0.453	0.605	0.409	0.421	0.352	0.293
	DT1	0.609	0.029	0.013	0.440	0.508	0.140	0.042	0.005	0.398	0.514	0.001	0.627	0.264	0.496	0.399	0.527	0.565	0.497	0.420	0.325	0.259
	DT2	0.627	0.123	0.001	0.374	0.501	0.239	0.004	0.002	0.482	0.356	0.004	0.569	0.331	0.441	0.248	0.494	0.616	0.456	0.415	0.329	0.288
	Norm	0.551	0.101	0.014	0.381	0.531	0.253	0.006	0.020	0.556	0.709	0.004	0.591	0.304	0.424	0.247	0.505	0.602	0.429	0.359	0.333	0.248
	Sdt	0.551	0.101	0.014	0.381	0.531	0.253	0.006	0.020	0.556	0.709	0.004	0.591	0.304	0.424	0.247	0.505	0.602	0.429	0.359	0.333	0.248
	Deriv1	0.638	0.129	0.005	0.390	0.461	0.225	0.039	0.206	0.344	0.575	0.006	0.591	0.299	0.420	0.207	0.487	0.621	0.439	0.433	0.361	0.281
	Deriv2	0.664	0.196	0.009	0.387	0.485	0.301	0.002	0.216	0.448	0.398	0.006	0.612	0.318	0.504	0.214	0.505	0.628	0.473	0.412	0.316	0.301
	MWA5	0.570	0.077	0.021	0.396	0.537	0.174	0.020	0.019	0.566	0.701	0.002	0.585	0.305	0.365	0.297	0.493	0.614	0.415	0.367	0.323	0.220
	MWA10	0.554	0.132	0.015	0.390	0.515	0.197	0.027	0.007	0.321	0.694	0.003	0.580	0.311	0.440	0.295	0.489	0.639	0.437	0.365	0.332	0.312
	MWA11	0.571	0.126	0.008	0.409	0.530	0.291	0.021	0.043	0.476	0.687	0.002	0.586	0.302	0.423	0.283	0.431	0.633	0.429	0.362	0.323	0.289
	MWA15	0.623	0.131	0.005	0.394	0.522	0.180	0.003	0.018	0.525	0.688	0.005	0.586	0.311	0.421	0.267	0.499	0.639	0.449	0.349	0.332	0.304
	MWA20	0.641	0.175	0.011	0.388	0.514	0.254	0.002	0.023	0.364	0.699	0.004	0.591	0.317	0.335	0.256	0.495	0.636	0.424	0.348	0.326	0.335
	CR	0.632	0.234	0.001	0.404	0.575	0.387	0.020	0.059	0.520	0.503	0.001	0.604	0.270	0.458	0.149	0.509	0.624	0.415	0.374	0.371	0.331

Color Scale
Minimum
Midpoint
Maximum

DISCUSSION

Model Selection

With a greater portion of the phenolic compound accumulation data, the random forest modeling techniques resulted in more accurate predictions and higher quality models, although there were a few cases where PLSR performed better. Both model types performed poorly when the data resulted in very low values, or the phenolic compound had a lot of missing values due to limited detection. These results may have been different with different data types or different sample sizes. For this specific set of data, the use of random forest models provided the higher quality models and better predictions. For similar datasets and types, random forest modeling should be explored as a high-quality way to predict phenotypes.

Although the random forest modeling technique performed better for a lot of the phenolic compounds and the preprocessing techniques, there are some constraints involved with this type

of model. The computational resources required to complete the random forest modeling for all 25 different preprocessing techniques and all 21 phenolic compounds is significantly greater than completing PLSR. To loop through all preprocessing techniques and phenolic compounds, the computational resources had to be moved to Michigan State University's high performance computing cluster (HPCC). On the HPCC, the script ran for over 26 hours and used 4CPUs and 100Gb of storage per CPU. In contrast, the loop for the PLSR modeling technique was able to run on a personal computer and was completed in less than an hour. The computational resources needed for the two different model types may play a role into model selection for future research projects. If there is a limitation on computational resources, the fine tuning of a PLSR model may be a better choice. If there is no limitation, or the researcher has access to an HPCC, a random forest model could provide the more accurate and high-quality choice.

Preprocessing techniques for spectral response data to aid in analysis quality

In contrast to the stand out success of random forest modeling, there was not an overall clear winner of the preprocessing methods. Each phenolic compound performed differently for each preprocessing technique. The only time the raw spectra outperformed the other preprocessing techniques was with apigenin-7-O-glucoside, but none of the models for this phenolic compound performed well. This aids in the understanding that raw spectrum does not perform well with modeling and should go through some type of preprocessing technique to remove noise and machine error and to perform better for modeling predictions. Overall, the preprocessing techniques that used the resampled spectra provided higher quality models for 14 of the phenolic compounds, in contrast to the 7 phenolic compounds that performed better with the techniques that used the raw spectra.

Standard normal variate and multiplicative scatter correction are two techniques that use very similar smoothing algorithms to remove the noise from the spectra. Multiplicative scatter correction is less prone to retaining spectral noise, but to use this method a reference spectrum is needed. The reference spectrum used in this case is the average of all the wavenumbers and not an outside spectrum. The standard normal variate technique differs because there is no need to provide the reference spectrum, but it is more prone to the noise. The use of a better reference spectrum with less error may increase the quality of the models created with the multiplicative scatter correction method.

Continuum removal was the best performing preprocessing method for one of the phenolic compounds. Normally, continuum removal is best used on small portions of the spectra and not the entire distance (Wardoux et al., 2021). Applying this technique to only wavenumbers on the spectra that coincide with the absorption of phenolic compounds could increase the success of this preprocessing technique.

The moving window average techniques with the different window sizes performed very similar to each other. The smoothing function of this method performed subtle adjustments to the spectra, but did not aid in huge differences in the result. A visual of the subtle adjustments can be seen in figure 4.3, with the color of each line representing a different window size. The use of multiple window sizes may not provide significantly different enough results to account for the computational resources required to test many different window sizes.

Overall, the testing of numerous different preprocessing methods for each of the phenolic compound showed that there is not a single method that stands out above the rest for all of the phenolic compounds. The testing of the different methods to create a different optimum model

for each phenolic compound is essential for getting the best prediction quality for using the spectral response data for phenotype prediction.

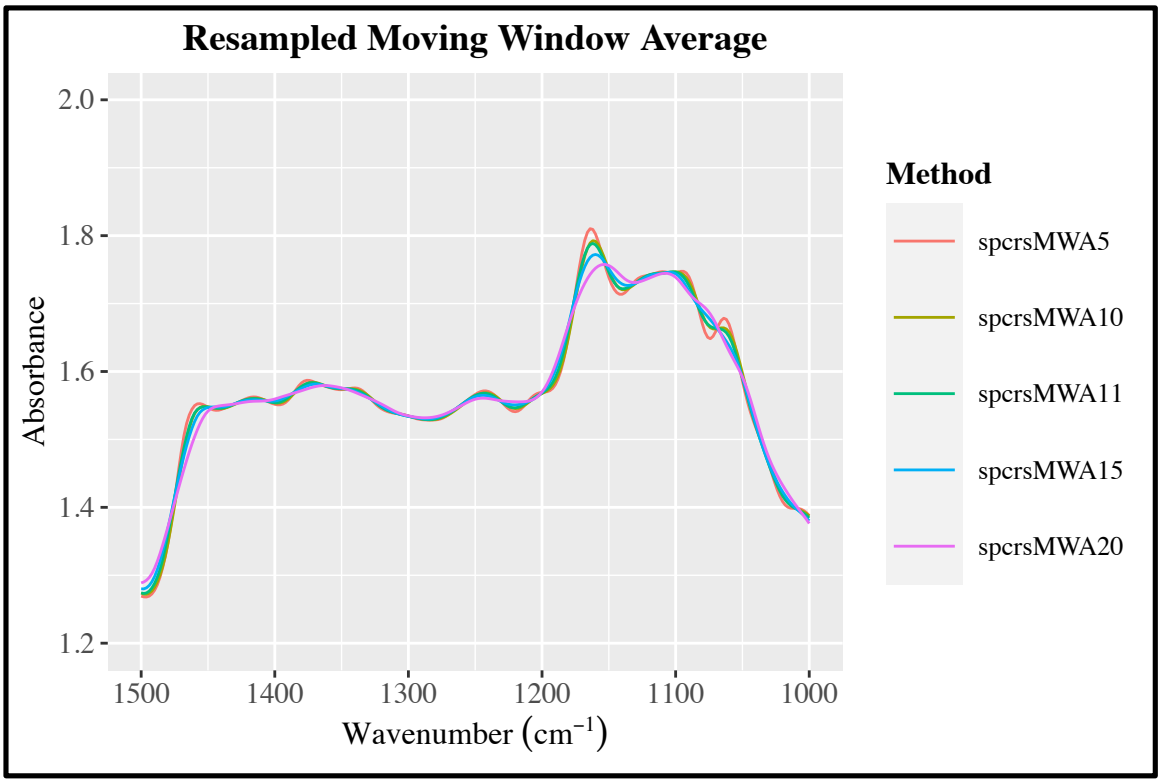


Figure 4.3: Plotted spectral response of the moving window average preprocessing techniques with window sizes of 5, 10, 11, 15, and 20 cm^{-1}. Transformations of the spectra were done on the resampled spectra and plot only includes the wavenumbers 1500 through 1000 cm^{-1} to highlight the subtle differences between the smoothing function. The line color indicates the different window sizes. Figure created using the R package ggplot2 (Wickham, 2016).

The use of spectral data for phenotyping

One of the main goals of this research study was to determine if there was a way to use spectral response data to predict phenolic compound accumulation data without the costly and time-consuming nature of the wet lab chemistry. The results of this study show promise for many of the phenolic compounds. Some of the R^2 values are above 0.7 and correlation between the observed and predicted data are above 0.84. Although there is still room for improvement, these results show success with models predicting phenotypes well. The addition of larger sample sizes

and greater diversity in phenolic compound accumulation phenotypes could provide the needed data for increasing the accuracy of the models even more than they currently are. Improving the models further may provide stand out models for each phenolic compound that could remove the need to spend excess time and money on the wet lab chemistry previously needed. Using FT-MIR spectroscopy could become an accurate method for phenotyping that is more accessible to researchers that do not have the time, funds, and facilities for the phenolic compound extraction and quantification through LC-MS.

Utilizing spectral data to aid in future plant breeding decisions.

Successful creation of models used to predict phenolic compound accumulation in plant tissues provides a more accessible way to include these phenotypes into plant breeding pipelines. Phenolic compounds are a very important element of plant defense and nutrition that are of interest to breeders. Producing these models would allow plant breeders to determine the phenolic compound content of their plant tissues with only spectral response data, which would remove the constraints provided by the laborious wet lab chemistry methods. Increasing the availability of phenolic compound accumulation data for plant breeders would allow additional progress to be made in plant breeding pipelines. This data could provide additional selection guidelines for breeders who want to explore increased nutritional values, plant stress response, and the plant immune system. Although, before these spectral techniques and models can be implemented, additional samples must be added to get a better view of the diversity of phenolic compound accumulation in maize. In addition to the use of kernels, these techniques could be applied to other plant tissues, such as a model for plant leaves or roots. The promise of the

research done in this study provides encouraging incentive to test these methods on other tissue types and with more phenolic compounds.

CONCLUSION

In conclusion, the research completed in this study shows promise with the ability to use FT-IR spectroscopy to model phenolic compound accumulation in maize kernels. Even though there are encouraging results, this study is limited based on the small sample sizes used to train and test the models, increasing the size could be beneficial to providing higher quality models and results. Also, this research showed the best results with using random forest models, but there was not a clear best choice for preprocessing method. To produce the most accurate phenotypes, individual models should be made for each phenolic compound that utilize the preprocessing method that provides the best result. This research aims to provide initial exploration of using spectroscopy to phenotype phenolic compound accumulation through the testing of numerous models and spectral processing techniques. This research also aims to encourage the additional investigation of this method of predicting phenotypes and to translate the methods on different tissue types. The continued advancements of FT-IR spectroscopy to model phenolic compound accumulation will provide accessible methods for plant breeders to use this data as a means of selection while limiting the costly and time-consuming wet lab chemistry.

REFERENCES

Abdi, H. (2003). Partial least square regression (PLS regression). *Encyclopedia for research methods for the social sciences*, *6*(4), 792-795.

Bachhar, V., Joshi, V., Gangal, A., Duseja, M., & Shukla, R. K. (2023). Identification of bioactive phytoconstituents, nutritional composition and antioxidant activity of calyptocarpus vialis. *Applied Biochemistry and Biotechnology*. https://doi.org/10.1007/s12010-023-04640-5

Berna, F. (2017). Fourier Transform Infrared Spectroscopy (FTIR). In: Gilbert, A.S. (eds) Encyclopedia of Geoarchaeology. Encyclopedia of Earth Sciences Series. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-4409-0_15

Borchers, H. W. (2022). pracma: Practical Numerical Math Functions. R package version 2.4.2, https://CRAN.R-project.org/package=pracma

Breiman L. (2001). Random forests. *Machine Learning* 45: 5–32

Bruker. (2023). *Guide to FT-IR spectroscopy*. Spectroscopy Basics. https://www.bruker.com/en/products-and-solutions/infrared-and-raman/ft-ir-routine-spectrometer/what-is-ft-ir-spectroscopy.html

Dutta, A. (2017). Fourier transform infrared spectroscopy. *Spectroscopic Methods for Nanomaterials Characterization*, 73–93. https://doi.org/10.1016/b978-0-323-46140-5.00004-2

Cocuron, J. C., Casas, M. I., Yang, F., Grotewold, E., & Alonso, A. P. (2019). Beyond the wall: High-throughput quantification of plant soluble and cell-wall bound phenolics by liquid chromatography tandem mass spectrometry. *Journal of Chromatography A*, *1589*, 93-104. https://doi.org/10.1016/j.chroma.2018.12.059

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: The next challenge. *Nature Reviews Genetics*, *11*(12), 855–866. https://doi.org/10.1038/nrg2897

JASCO Inc. (2023, May 17). *FTIR spectroscopy (overview)*. Fundamental Theory and Applications of FTIR Spectroscopy. https://jascoinc.com/learning-center/theory/spectroscopy/fundamentals-ftir-spectroscopy/#FTIR-spectroscopy-principles

Kalendar, R., Ghamkhar, K., Franceschi, P., & Egea-Cortines, M. (2022). Editorial: Spectroscopy for crop and product phenotyping. Frontiers in Plant Science, 13. https://doi.org/10.3389/fpls.2022.1058333

Kumar, J., Pratap, A., & Kumar, S. (2015). Plant Phenomics: An Overview. In: Kumar, J., Pratap, A., Kumar, S. (eds) Phenomics in Crop Plants: Trends, Options and Limitations. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2226-2_1

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. R News *2*(3), 18-22. https://CRAN.R-project.org/doc/Rnews/

Liland K., Mevik, B., & Wehrens, R. (2023). pls: Partial Least Squares and Principal Component Regression. R package version 2.8-2, https://CRAN.R-project.org/package=pls

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems, 118*, 62-69.

Mierzwa-Hersztek, M., Gondek, K., Mierzwa-Hersztek, M., Nawrocka, A., Pińkowska, H., Bajda, T., Stanek-Tarkowska, J., & Szostek, M. (2019). FT-IR analysis and the content of phenolic compounds in exogenous organic matter produced from plant biomass. *Journal of Elementology*, (3/2019). https://doi.org/10.5601/jelem.2018.23.3.1716

Núñez, E., Steyerberg, E. W., & Núñez, J. (2011). Regression modeling strategies. *Revista Española de Cardiología (English Edition)*, *64*(6), 501–507. https://doi.org/10.1016/j.rec.2011.01.017

Rodriguez, J., Gomez-Cano, L., Grotewold, E., & de Leon, N. (2022). Normalizing and correcting variable and complex LC–ms metabolomic data with the R package pseudodrift. *Metabolites*, *12*(5), 435. https://doi.org/10.3390/metabo12050435

Schonlau, M., & Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867x20909688

Signal Developers (2014). *signal: Signal processing*. http://r-forge.r-project.org/projects/signal/

Stevens A. & Ramirez-Lopez, L. (2022). *An introduction to the prospectr package*. R package version 0.2.6.

Stevenson, F.J. 1982 *Humus Chemistry: Genesis, Composition, Reactions*. John Wiley & Sons, New York.

Wadoux, A. M. J.-C., Malone, B., Minasny, B., Fajardo, M., & McBratney, A. B. (2021). *Soil Spectra Inference with R: Analyzing Digital Soil Spectra using the R Programming Environment* (A. E. Hartemink & A. B. McBratney, Eds.; Ser. Progress in Soil Science). Springer.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. *Springer-Verlag New York.* ISBN: 978-3-319-24277-4, https://ggplot2.tidyverse.org