GENETIC AND MOLECULAR DISSECTION OF BIOMASS COMPOSITIONAL TRAITS IN MAIZE

by

Jonas Rodriguez

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Plant Breeding and Plant Genetics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: 4/19/2022

This dissertation is approved by the following members of the Final Oral Committee:
Natalia de Leon, Professor, Agronomy
Edgar P. Spalding, Professor, Botany
Hiroshi Maeda, Associate Professor, Botany
Shawn M. Kaeppler, Professor, Agronomy
Steven D. Karlen, Scientist, Wisconsin Energy Institute

DEDICATION

A mi querida mamá, Catalina y mis hermanos, Judy, Junior, Amy, & Daisy

ACKNOWLEDGEMENTS

This thesis would not have been possible without the unwavering support from many people along the way. First and foremost, I would like to thank my advisor and mentor Dr. Natalia de Leon. Her kindness, optimism, and genuine interest in seeing others succeed is unparalleled and I am incredibly thankful to her for providing me with the opportunity to join her group. I would also like to thank Dr. Shawn Kaeppler for his thoughtfulness and guidance. I am also grateful to the other members of my committee, Dr. Edgar Spalding, Dr. Hiroshi Maeda, and Dr. Steven Karlen, for their valuable input to this thesis. I would like to acknowledge Dr. Erich Grotewold and Lina Gomez-Cano, for their contributions to this thesis and for being nothing short of wonderful collaborators. I would have never pursued graduate school if it were not for Dr. Julie Ho, Dr. Charlie Brummer, and Dr. Eduardo Blumwald. I am forever thankful to them for their sound advice and encouragement. Also, I owe thanks to the entire field corn group, current and former members, for managing to make long pollen drenched days in the field oddly enjoyable.

I cannot understate the gratitude I have towards the friends I have made during my time as a graduate student. My time in Wisconsin would not have been nearly as enjoyable without you all. Lastly, I would like to give a special thanks to my wife, Michelle Hrdi. Thank you for providing unconditional love, and reminding me to always stay true to my best self.

ABSTRACT

In the context of converting plant biomass to derived value added products, biomass composition is a major determinant of biomass utility. Maize (*Zea mays* L.) is one of the most widely produced and utilized crops, and the biomass most sought after is the grain-portion. However, the importance of the non-grain (vegetative) portion of the biomass should not be understated. This thesis investigates several aspects of vegetative biomass composition using genetic and analytical chemistry tools and techniques. The first chapter provides an overview of maize biomass composition and introduces the approaches used in the second, third and fourth chapters to study compositional traits. The second chapter reports a genome-wide association and gene co-expression analysis to identify novel candidate genes associated with stalk anatomical and saccharification efficiency traits. A total of 62 candidate genes were prioritized based on these genetic analyses. The third chapter reports a new statistical approach and implementation tool for identifying and adjusting for technical errors commonly encountered in metabolomics data. The utility of the tool is evaluated using simulated data, and the statistical approach is applied to a newly generated large-scale maize phenolic metabolite dataset. The last chapter uses data generated from the third chapter and reports a genome and transcriptome-wide association analysis, along with a differential gene expression analysis to identify candidate genes associated with the accumulation of four phenolic compounds. A total of 20 top candidate genes are suggested for guiding future validation studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY FILES

Chapter 1: Literature Review

1.1 Maize biomass and composition

Overview of maize biomass

Plant biomass is defined as the total above and below-ground organic matter that a plant produces through photosynthesis converting light to chemical energy that the plant either uses or stores (Johnson, 2016; Roberts et al., 1985). In mature maize plants, the above-ground plant tissues represent the largest proportion of biomass and include the stalk, leaves, cob, grain, and tassel. The caloric (energy) distribution of the above-ground biomass varies by tissue, with the grain accounting for approximately half of the total energy and the remainder corresponding to the stover (e.g., cob, leaves, stalk, tassel) (Hedin et al., 1998). Biochemically, the composition of different plant tissues primarily explains the observed caloric distribution among maize tissues. Densely packed starch molecules make up a substantial proportion of the grain, while the stover consists mainly of lignocellulosic (cellulose, hemicellulose, and lignin) biopolymers of cell wall material (Boyer & Shannon, 2003; Hansey et al., 2010).

Stover versus grain biomass

Unsurprisingly, maize cultivation is primarily for grain due to its high energy density. However, stover and grain are tightly linked, a relationship termed harvest index. For example, Lorenz et al. (2010) re-analyzed data from five studies (Lauer et al., 2001; Meghji et al., 1984; Russel, 1984; Tapper, 1983; Tollenaar, 1989) that evaluated Corn Belt maize hybrids from different eras planted at low to intermediate densities, and found proportional increases in hybrid grain and stover yield per unit area over several decades. Their results conflicted with an earlier study that also evaluated Corn Belt maize hybrids from different eras (Duvick et al., 2003), which reported increases in grain yield but not stover yield on a per plant basis. That is, Duvick et al. (2003), suggested that maize breeders had successfully increased the ratio of grain yield to total biomass (harvest index) over time. However, Lorenz et al. (2010) noted that in Duvick et al. (2003), the authors had evaluated hybrids from different eras uniformly, but under high planting densities that favored grain production of newer hybrids selectively bred to tolerate high planting densities. In contrast, the older hybrids did not have the same tolerance.

When planted at high densities, maize hybrids (new and old alike) experience a significant amount of stress due to competition for resources with neighboring plants (Tollenaar et al., 1994). As a response, plants tend to have narrower and weaker stalks (Stanger & Lauer, 2007), with presumably less stover biomass per plant. Additionally, at extreme densities ears can become barren and produce no grain at all (Tollenaar et al., 1994). Thus, Lorenz et al. (2010) hypothesized that the increases in harvest index reported by Duvick et al. (2003) would likely not be observed if instead lower planting densities were used. The assumption was that there would be negligible competition stress, and therefore newer hybrids might still produce more grain but would produce more stover as well.

Utility of maize biomass

A net-zero change over time in the harvest index of maize is not universally a negative attribute. Unlike maize cultivated for its grain, when maize is used as a forage for animal feed (silage), the total biomass is a valuable asset at harvest. Silage is usually harvested a few weeks before grain physiological maturity, when whole plant moisture drops to between 60 - 70% (Wiersma et al., 1993). Harvesting outside of this range can cause potential problems when ensiling and is generally associated with lower feed quality (Wiersma et al., 1993). Ultimately, feed quality is determined by animal performance, such as milk production by dairy cows (National Research Council, 2001). The most common evaluation methods of silage quality are based on a series of detergent (Soest & Wine, 1967) and *in vitro* (Tilley & Terry, 1963) assays which provide estimates of biomass digestibility by ruminants. For example, neutral detergent fiber digested (NDFD) represents the portion of lignocellulosic material digested by an animal and is a key input to predicting milk production potential from silage (National Research Council, 2001).

When cultivated for grain production, the stover portion of the biomass receives minimal attention after physiological maturity since sugars are no longer being translocated to the kernels (Daynard & Duncan, 1969). At the end of the growing season, the stover is largely seen as merely a crop residue. However, stover biomass and quality can be important when used as a lignocellulosic feedstock by biorefineries. Stover and silage quality improvement are similar in that both are uses that convert biomass into a value added product (Lorenz & Coors, 2008). Silage is used in milk production, whereas lignocellulosic biomass can be used by biorefineries to produce fuels or other value added products (Upton & Kasko, 2016; Valdivia et al., 2016). With established breeding methods and approaches for silage improvement, efforts have been made to leverage this knowledge to improve stover composition for bioconversion (Lorenz & Coors, 2008). For example, by subjecting a synthetic population of maize developed for silage to four cycles of reciprocal selection for increased total biomass and feed quality, Gustafson et al. (2010) observed a concurrent increase in feed quality, whole-plant yield, and stover yield at silage maturity both on a *per se* as well as test-cross basis. However, at grain physiological maturity, stover yield and quality, both on a *per se* and test-cross basis, did not change between cycles of selection. Thus, the observations by Gustafson et al. (2010) underscore the distinction between silage and stover improvement and the potential for stover-specific selection in maize.

Biosynthesis and structure of lignocellulosic components

The composition of maize stover on a dry matter basis is comprised primarily of the polysaccharides cellulose (~35%) and hemicellulose (~20%) and, to a lesser extent, the recalcitrant plant phenolic polymer, lignin (~12%) (Ruan et al., 2019). The remaining ~33% of the mass corresponds to soluble proteins, starch, and minerals.

The structure of cellulose is conserved in all plant species, and it is organized into ~25-30 nm diameter microfibrils that are made up of long chains of β-1,4 linked glucose molecules called glucans (Somerville, 2006). Cellulose biosynthesis occurs at the plasma membrane by cellulose synthase (CESA) complexes, which are comprised of several CESA peptides. At the membrane, each CESA peptide synthesizes an individual glucan molecule, and on the other side of the membrane, glucans derived from the same CESA complex are joined to form a microfibril (Delmer & Amor, 1995; Mazur & Zimmer, 2011; Pear et al., 1996). Despite the characterization of the protein complexes involved, the number of glucans making up a microfibril remains to be definitively proven. Early work suggested a 36-chain model (Herth, 1983)**,** and similar ideas persisted based on microscopic observations of six-fold symmetrical rosette structures formed by CESA complexes. However, several recently published studies have used advanced techniques, including crystallography, solid-state nuclear magnetic resonance (NMR), and computational modeling approaches, that offer the most compelling evidence yet to support an 18-chain model instead (Morgan et al., 2013; Newman et al., 2013; Nixon et al., 2016; Sethaphong et al., 2013).

Namely, that a 6 x 3 CESA peptide assembly as a trimer is consistent with physical size and rosette shape characteristics of CESA complexes.

Unlike cellulose, which is a homopolymer of glucose, hemicellulose refers to several non-cellulose polysaccharides (Pauly et al., 2013). The diversity of hemicelluloses is exemplified by the fact that for some families of flowering plants the hemicellulose composition and structure alone can distinguish one family from another (Dahlgren, 1989; The Angiosperm Phylogeny Group et al., 2016). The most prevalent hemicellulose in maize and most angiosperms is xyloglucan (Hsieh & Harris, 2009), which is biosynthesized in the Golgi and subsequently excreted to the cell wall by transport vesicles (Drakakaki & Dandekar, 2013). The steps involved in xyloglucan biosynthesis include first building the β-1,4 linked glucosyl backbone, followed by several substitutions with other sugars such as galactose and fucose to generate a branched polysaccharide (Pauly & Keegstra, 2016).

Lignin structure is notoriously complex, and current research is altering the notion of what typifies its structure. The canonical components of lignin are three lignin monomers (monolignols) that include p-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol, which are synthesized in the cytosol by a branch of the phenylpropanoid pathway (Vermerris, 2008). Monolignols are eventually transported to the cell wall, where they undergo oxidation by peroxidases and subsequently, through a purely chemical process, polymerize by oxidative radical coupling (Ralph et al., 2004, 2008). In addition to the three classic monolignols, recent studies have observed non-conventional phenolics incorporated into the lignin polymer of several plant species (del Río et al., 2020, 2022; Karlen et al., 2016, 2017; Lu et al., 2015). Through these observations, the authors have reemphasized the complexity of lignin composition and the evolving definition of what constitutes the polymer.

1.2 Methods for evaluating lignocellulosic biomass saccharification

Fermentable sugar assays

The objective of lignocellulosic biomass saccharification is to access the simple monomeric sugars that make up the cellulose and hemicellulose polysaccharides, namely glucose (GLU) and pentose (PEN). This process is most hampered by the presence of lignin embedded in the cell wall matrix. There are various techniques to promote the efficient saccharification of biomass, including different combinations of mechanical, thermal and chemical pretreatments, and the addition of exogenous polysaccharide degrading enzymes (Khare et al., 2015). Additionally, the composition of feedstocks can have a major impact on saccharification efficiency (Bichot et al., 2018). Therefore, substantial efforts have been placed into developing high-throughput and pre-treatment flexible saccharification assay platforms (Chundawat et al., 2008; Decker et al., 2009; L. D. Gomez et al., 2010; Patil et al., 2022; Studer et al., 2010). The iWALL platform (Santoro et al., 2010), in particular, has enabled recent large-scale genetic studies related to biomass recalcitrance in Arabidopsis (Sakamoto et al., 2020) switchgrass (Saha et al., 2021), and maize (Kumar et al., 2021).

Potential need for alternative assays

Current biorefineries operate under a polysaccharide-first paradigm as described above. However, a lignin-first alternative biomass processing paradigm has recently been proposed (Abu-Omar et al., 2021; Korányi et al., 2020). Lignin-first biorefineries would place an emphasis on preserving the valuable lignin-derived phenolic compounds (e.g., ferulic acid and *p*-coumaric acid) during the polysaccharide extraction processing steps. This would enable a biorefinery to diversify its product outputs (Upton & Kasko, 2016) and, in turn, provide increased economic buffering and viability to an industry that has faced substantial challenges especially when oil prices have been low. If lignin-first biorefining were to increase in prevalence, new high-throughput biomass evaluation methods or modifications to existing platforms might be needed.

1.3 Maize stalk anatomical characteristics associated with biomass saccharification

Stalk anatomical structures

The anatomical structures of mature maize stalks are paradoxically simple and complex. Visually, their complexity is best illustrated at the nodes where vascular bundles form highly interwoven networks (Cheng et al., 2001; Shane, 2000). In contrast, in the internodes, the pith, rind, and individual pith-localized vascular bundles are easily distinguished with a naked eye. Although visually simple, the internode is compositionally and functionally complex. For example, the fundamentally important process of nutrient and water mobilization occurs through the xylem and phloem of the vascular bundles, both those in the pith and embedded in the rind (Crang et al., 2018; Shane, 2000). Compositionally, the degree of lignification differs between the three main structures, with the rind and vascular bundles being the most lignified and resistant to degradation (Wilson & Hatfield, 1997). Therefore, anatomical characteristics have been suggested as possible major determinants of cell wall degradability (Wilson & Hatfield, 1997; Wilson & Mertens, 1995). While visually dissecting and manually quantifying anatomical features on a few dissected maize internodes might be straightforward, a major bottleneck involves scaling up to evaluate hundreds or thousands of plant samples.

Phenotyping stalk anatomical structures

Due to the essential role that the vasculature system serves in plants, Strock et al., (2022) recently coined the term Anatomics to describe high-throughput plant anatomy phenotyping. Several methods have emerged specifically for extracting anatomical trait information from stalk internodes, including image-based (Heckwolf et al., 2015), x-ray (F. E. Gomez et al., 2018; Zhang et al., 2018), and computed tomography (CT) based (Zhang et al., 2020) approaches. These phenotyping methods have different levels of resolution and throughput. For example, the image-based method from Heckwolf et al. (2015) offers the highest throughput and quantifies the characteristics of pith-localized vascular bundles but does not provide metrics for rind-localized vascular bundles like the method developed by Zhang et al. (2020). Recently, the method from Heckwolf et al. (2015) was applied to evaluate and genetically dissect several anatomical characteristics from a population of 942 diverse maize inbred lines (Mazaheri et al., 2019). Thus, demonstrating the scalability and utility of the method.

1.4 Molecular phenotypes and associated challenges

Intermediates between genotype and phenotype

Modern sequencing, analytical techniques, and instrumentation are rapidly advancing and enabling a more comprehensive understanding of the steps from genotype to phenotype. Genetic analysis is now so ubiquitous that human genetic testing is marketed directly to consumers, and the names of the companies offering these services have essentially become eponyms (Su, 2013). As of January 2nd, 2022, the Sequence Read Archive (SRA) holds over 12 petabytes of open access data (International Nucleotide Sequence Database Collaboration, 2022), which is a testament to how common next-generation sequence data generation is among researchers.

RNA-Seq has become the defacto genome-wide transcriptional profiling technology, which can also be used for single nucleotide polymorphism (SNP) discovery (Zhao et al., 2019). There are many large publicly available maize RNA-Seq datasets that capture transcriptional variation through development in different tissues, genetic variation, and genotype x tissue variability (Chen et al., 2014; Eichten et al., 2013; Fu et al., 2013; Hirsch et al., 2014; Hoopes et al., 2019; Huang et al., 2018; Kremling et al., 2018; Leiboff et al., 2015; Li et al., 2019; Lin et al., 2017; Mazaheri et al., 2019; Schaefer et al., 2018; Sekhon et al., 2011; Stelpflug et al., 2016; Walley et al., 2016). Gene expression data generated from these publicly available sequencing reads have been used to conduct a meta-analysis of gene regulatory interactions (P. Zhou et al., 2020), which would not have been possible without these data as a public resource. Although gene expression can be useful to conduct different types of analyses to answer or generate new hypotheses, transcript abundance is not always highly correlated with protein abundance and does not represent true intermediates like metabolites that give rise to phenotypes (Nachtomy et al., 2007).

In recent years there has been a growing interest in generating and using molecular intermediates such as metabolites due to their direct effects on phenotypes. For example, the canonical and non-canonical monolignols that form the lignin polymer are all metabolites from the phenylpropanoid pathway (Ralph et al., 2004, 2008; Vermerris, 2008). Flavonoids, which are also produced by the phenylpropanoid pathway, have also been extensively characterized primarily for their contribution to pigmentation phenotypes (Grotewold, 2006). However, there are still many unanswered questions related to gene regulation, transport, and other specialized functions that may be crucial to understanding the mechanisms of metabolism (Biała & Jasiński, 2018; Dixon et al., 2002; Gray et al., 2012; Sharma et al., 2019). A recent study developed a targeted liquid chromatography-mass spectrometry (LC–MS) method to profile approximately thirty phenolic compounds, including many flavonoids, from the phenylpropanoid pathway, which they validated using stem material extracts from a set of three genetically distinct maize seedlings (Cocuron et al., 2019). The compounds targeted by this method represent a core set of phenolics that give rise to the several thousand phenolic compounds produced by plants. Thus, this method and potentially other methods can be applied to larger and more diverse populations to facilitate the investigation of unknown processes occurring within the phenylpropanoid pathway.

Technical challenges associated with metabolic profiling

Despite recent advances in analytical techniques that allow for high-resolution metabolic profiling (Lelli et al., 2021), the scaling up of these techniques can introduce several challenges. In targeted LC-MS, for example, depending on the number of samples being evaluated and methods utilized, run times can quickly add up to span several days or potentially weeks. While long run times might not be an issue for groups with their own instrumentation, it can be limiting for those who rely on shared resources. A tangible solution then might be to split large runs into several batches. However, partitioning runs can result in technical errors manifested as batch effects, which are widely recognized, but not easily addressed since each batch might contain a unique set of variables contributing to the observed error (Wehrens et al., 2016). Furthermore, another common source of technical error comes from gradual changes in the instrumentation response within a batch known as ionization drift, or simply signal drift (Kuligowski et al., 2015). Consequently, essentially all large-scale LC-MS experiments include a post-data acquisition statistical analysis aimed at reducing technical errors.

Several different statistical methods exist to correct for signal drift and batch effects in LC-MS data (Fernández-Albert et al., 2014). Rather than explicitly detailing each individually, a summary of the two classes of methods is described below.

The first class is dependent on the inclusion of quality control (QC) samples in the run that are interspersed at pre-determined intervals between experimental samples to be analyzed throughout the LC-MS run. QC samples are typically derived from an identically prepared extract, which enables for a statistical correction to be applied based on the assumption that identical QC samples should return identical values. The second class of methods does not rely on a QC but instead assumes that variances are equal within and between batches. Under this assumption, a statistical correction is then applied based on a parameter derived from the data, such as the mean or median, of the instrument response within batches and then between batches.

Generally, the consensus is that QC inclusion followed by a QC-based correction method provides the best performance in reducing technical errors (Broadhurst et al., 2018). However, a similar consensus does not exist on what constitutes a ‘good’ QC or what the optimal QC frequency should be within a run. Therefore, guidelines that provide recommendations for one field of study may not be ideal for other fields. For instance, urine is commonly analyzed with untargeted LC-MS by clinical researchers and has been described as a biologically complex matrix containing several hundred metabolites (Rodríguez-Morató et al., 2018; Wang et al., 2022). In comparison, a recent maize study also analyzed by untargeted evaluated leaf bases and tips and identified nearly 4,000 metabolic features (S. Zhou et al., 2019). Given these rather dramatic differences in sheer metabolite diversity, it is unclear whether both would benefit from utilizing the same experimental guidelines or downstream correction techniques.

With the recent creation of cross-species metabolomics centered databases such as Metabolights (Haug et al., 2020), it is possible that species or tissue-specific recommendations might emerge. Even if specific guidelines are not established, the data (and meta-data) represent a rich resource for the many researchers planning new LC-MS experiments or re-analyzing already published datasets.

1.5 References

Abu-Omar, M. M., Barta, K., Beckham, G. T., Luterbacher, J. S., Ralph, J., Rinaldi, R., Román-Leshkov, Y., Samec, J. S. M., Sels, B. F., & Wang, F. (2021). Guidelines for performing lignin-first biorefining. *Energy & Environmental Science*, *14*(1), 262–292. https://doi.org/10.1039/D0EE02870C

Biała, W., & Jasiński, M. (2018). The Phenylpropanoid Case – It Is Transport That Matters. *Frontiers in Plant Science*, *9*, 1610. https://doi.org/10.3389/fpls.2018.01610

Bichot, A., Delgenès, J.-P., Méchin, V., Carrère, H., Bernet, N., & García-Bernet, D. (2018). Understanding biomass recalcitrance in grasses for their efficient utilization as biorefinery feedstock. *Reviews in Environmental Science and Bio/Technology*, *17*(4), 707–748. https://doi.org/10.1007/s11157-018-9485-y

Boyer, C. D., & Shannon, J. C. (2003). Carbohydrates of the kernel. *Corn: Chemistry and Technology, Ed.2*, 289–311.

Broadhurst, D., Goodacre, R., Reinke, S. N., Kuligowski, J., Wilson, I. D., Lewis, M. R., & Dunn, W. B. (2018). Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, *14*(6), 72. https://doi.org/10.1007/s11306-018-1367-3

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A., & Lai, J. (2014). Dynamic Transcriptome Landscape of Maize Embryo and Endosperm Development. *Plant Physiology*, *166*(1), 252–264. https://doi.org/10.1104/pp.114.240689

Cheng, P. C., Chen, J. H., Hwang, S. C., Sun, C. K., Walden, D. B., & Cheng, W. Y. (2001). 3D Visualization of Maize Stem by MRI Technology. *Microscopy and Microanalysis*, *7*(S2), 100–101. https://doi.org/10.1017/S143192760002657X

Chundawat, S. P. S., Balan, V., & Dale, B. E. (2008). High-throughput microplate technique for enzymatic hydrolysis of lignocellulosic biomass. *Biotechnology and Bioengineering*, *99*(6), 1281–1294. https://doi.org/10.1002/bit.21805

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

Crang, R., Lyons-Sobaski, S., & Wise, R. (2018). *Plant Anatomy: A Concept-Based Approach to the Structure of Seed Plants*. Springer International Publishing. https://doi.org/10.1007/978-3-319-77315-5

Dahlgren, G. (1989). An updated angiosperm classification. *Botanical Journal of the Linnean Society*, *100*(3), 197–203. https://doi.org/10.1111/j.1095-8339.1989.tb01717.x

Daynard, T. B., & Duncan, W. G. (1969). The Black Layer and Grain Maturity in Corn1. *Crop Science*, *9*(4), 473–476. https://doi.org/10.2135/cropsci1969.0011183X000900040026x

Decker, S. R., Brunecky, R., Tucker, M. P., Himmel, M. E., & Selig, M. J. (2009). High-Throughput Screening Techniques for Biomass Conversion. *BioEnergy Research*, *2*(4), 179. https://doi.org/10.1007/s12155-009-9051-0

del Río, J. C., Rencoret, J., Gutiérrez, A., Elder, T., Kim, H., & Ralph, J. (2020). Lignin Monomers from beyond the Canonical Monolignol Biosynthetic Pathway: Another Brick in the Wall. *ACS Sustainable Chemistry & Engineering*, *8*(13), 4997–5012. https://doi.org/10.1021/acssuschemeng.0c01109

del Río, J. C., Rencoret, J., Gutiérrez, A., Kim, H., & Ralph, J. (2022). Unconventional lignin monomers—Extension of the lignin paradigm. In *Advances in Botanical Research*. Academic Press. https://doi.org/10.1016/bs.abr.2022.02.001

Delmer, D. P., & Amor, Y. (1995). Cellulose biosynthesis. *The Plant Cell*, *7*(7), 987–1000.

Dixon, R. A., Achnine, L., Kota, P., Liu, C.-J., Reddy, M. S. S., & Wang, L. (2002). The phenylpropanoid pathway and plant defence—A genomics perspective. *Molecular Plant Pathology*, *3*(5), 371–390. https://doi.org/10.1046/j.1364-3703.2002.00131.x

Drakakaki, G., & Dandekar, A. (2013). Protein secretion: How many secretory routes does a plant cell have? *Plant Science*, *203–204*, 74–78. https://doi.org/10.1016/j.plantsci.2012.12.017

Duvick, D. N., Smith, J. S. C., & Cooper, M. (2003). Long-Term Selection in a Commercial Hybrid Maize Breeding Program. In *Plant Breeding Reviews* (pp. 109–151). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470650288.ch4

Eichten, S. R., Briskine, R., Song, J., Li, Q., Swanson-Wagner, R., Hermanson, P. J., Waters, A. J., Starr, E., West, P. T., Tiffin, P., Myers, C. L., Vaughn, M. W., & Springer, N. M. (2013). Epigenetic and Genetic Influences on DNA Methylation Variation in Maize Populations. *The Plant Cell*, *25*(8), 2783–2797. https://doi.org/10.1105/tpc.113.114793

Fernández-Albert, F., Llorach, R., Garcia-Aloy, M., Ziyatdinov, A., Andres-Lacueva, C., & Perera, A. (2014). Intensity drift removal in LC/MS metabolomics by common variance compensation. *Bioinformatics*, *30*(20), 2899–2905. https://doi.org/10.1093/bioinformatics/btu423

Fu, J., Cheng, Y., Linghu, J., Yang, X., Kang, L., Zhang, Z., Zhang, J., He, C., Du, X., Peng, Z., Wang, B., Zhai, L., Dai, C., Xu, J., Wang, W., Li, X., Zheng, J., Chen, L., Luo, L., … Wang, G. (2013). RNA sequencing reveals the complex regulatory network in the maize kernel. *Nature Communications*, *4*, 2832. https://doi.org/10.1038/ncomms3832

Gomez, F. E., Carvalho, G., Shi, F., Muliana, A. H., & Rooney, W. L. (2018). High throughput phenotyping of morpho-anatomical stem properties using X-ray computed tomography in sorghum. *Plant Methods*, *14*(1), 1–13. https://doi.org/10.1186/s13007-018-0326-3

Gomez, L. D., Whitehead, C., Barakate, A., Halpin, C., & McQueen-Mason, S. J. (2010). Automated saccharification assay for determination of digestibility in plant materials. *Biotechnology for Biofuels*, *3*, 23. https://doi.org/10.1186/1754-6834-3-23

Gray, J., Caparrós-Ruiz, D., & Grotewold, E. (2012). Grass phenylpropanoids: Regulate before using! *Plant Science*, *184*, 112–120. https://doi.org/10.1016/j.plantsci.2011.12.008

Grotewold, E. (2006). The Genetics and Biochemistry of Floral Pigments. *Annual Review of Plant Biology*, *57*(1), 761–780. https://doi.org/10.1146/annurev.arplant.57.032905.105248

Gustafson, T. J., Coors, J. G., & de Leon, N. (2010). Selection for forage yield and composition on the Wisconsin Quality Synthetic maize population. *Crop Science*, *50*(5), 1795–1804. https://doi.org/10.2135/cropsci2009.12.0725

Hansey, C. N., Lorenz, A. J., & de Leon, N. (2010). Cell Wall Composition and Ruminant Digestibility ofVarious Maize Tissues Across Development. *Bioenergy Research*, *3*(3), 295–304. https://doi.org/10.1007/s12155-010-9100-8

Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., & O’Donovan, C. (2020). MetaboLights: A resource evolving in response to the needs of its scientific community. *Nucleic Acids Research*, *48*(D1), D440–D444. https://doi.org/10.1093/nar/gkz1019

Heckwolf, S., Heckwolf, M., Kaeppler, S. M., de Leon, N., & Spalding, E. P. (2015). Image analysis of anatomical traits in stalk transections of maize and other grasses. *Plant Methods*, *11*(26), 26. https://doi.org/10.1186/s13007-015-0070-x

Hedin, P. A., Williams, W. P., & Buckley, P. M. (1998). Caloric Analyses of the Distribution of Energy in Corn Plants Zea mays L. *Journal of Agricultural and Food Chemistry*, *46*(11), 4754–4758. https://doi.org/10.1021/jf980439z

Herth, W. (1983). Arrays of plasma-membrane “rosettes” involved in cellulose microfibril formation of Spirogyra. *Planta*, *159*(4), 347–356. https://doi.org/10.1007/BF00393174

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell*, *26*(1), 121–135. https://doi.org/10.1105/tpc.113.119982

Hoopes, G. M., Hamilton, J. P., Wood, J. C., Esteban, E., Pasha, A., Vaillancourt, B., Provart, N. J., & Buell, C. R. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *The Plant Journal*, *97*(6), 1154–1167. https://doi.org/10.1111/tpj.14184

Hsieh, Y. S. Y., & Harris, P. J. (2009). Xyloglucans of Monocotyledons Have Diverse Structures. *Molecular Plant*, *2*(5), 943–965. https://doi.org/10.1093/mp/ssp061

Huang, J., Zheng, J., Yuan, H., & McGinnis, K. (2018). Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize. *BMC Plant Biology*, *18*(1), 1–14. https://doi.org/10.1186/s12870-018-1329-y

International Nucleotide Sequence Database Collaboration. (2022). https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?

Johnson, M. P. (2016). Photosynthesis. *Essays in Biochemistry*, *60*(3), 255–273. https://doi.org/10.1042/EBC20160016

Karlen, S. D., Smith, R. A., Kim, H., Padmakshan, D., Bartuce, A., Mobley, J. K., Free, H. C. A., Smith, B. G., Harris, P. J., & Ralph, J. (2017). Highly Decorated Lignins in Leaf Tissues of the Canary Island Date Palm Phoenix canariensis. *Plant Physiology*, *175*(3), 1058–1067. https://doi.org/10.1104/pp.17.01172

Karlen, S. D., Zhang, C., Peck, M. L., Smith, R. A., Padmakshan, D., Helmich, K. E., Free, H. C. A., Lee, S., Smith, B. G., Lu, F., Sedbrook, J. C., Sibout, R., Grabber, J. H., Runge, T. M., Mysore, K. S., Harris, P. J., Bartley, L. E., & Ralph, J. (2016). Monolignol ferulate conjugates are naturally incorporated into plant lignins. *Science Advances*, *2*(10), e1600393. https://doi.org/10.1126/sciadv.1600393

Khare, S. K., Pandey, A., & Larroche, C. (2015). Current perspectives in enzymatic saccharification of lignocellulosic biomass. *Biochemical Engineering Journal*, *102*, 38–44. https://doi.org/10.1016/j.bej.2015.02.033

Korányi, T. I., Fridrich, B., Pineda, A., & Barta, K. (2020). Development of 'Lignin-First' Approaches for the Valorization of Lignocellulosic Biomass. *Molecules*, *25*(12), 2815. https://doi.org/10.3390/molecules25122815

Kremling, K. A. G., Chen, S. Y., Su, M. H., Lepak, N. K., Romay, M. C., Swarts, K. L., Lu, F., Lorant, A., Bradbury, P. J., & Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, *555*(7697), 520–523. https://doi.org/10.1038/nature25966

Kuligowski, J., Sánchez-Illana, Á., Sanjuán-Herráez, D., Vento, M., & Quintás, G. (2015). Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC). *Analyst*, *140*(22), 7810–7817. https://doi.org/10.1039/C5AN01638J

Kumar, R., Gyawali, A., Morrison, G. D., Saski, C. A., Robertson, D. J., Cook, D. D., Tharayil, N., Schaefer, R. J., Beissinger, T. M., & Sekhon, R. S. (2021). Genetic Architecture of Maize Rind Strength Revealed by the Analysis of Divergently Selected Populations. *Plant & Cell Physiology*, *62*(7), 1199–1214. https://doi.org/10.1093/pcp/pcab059

Lauer, J. G., Coors, J. G., & Flannery, P. J. (2001). Forage yield and quality of corn cultivars developed in different eras. *Crop Science*, *41*(5), 1449–1455.

Leiboff, S., Li, X., Hu, H.-C., Todt, N., Yang, J., Li, X., Yu, X., Muehlbauer, G. J., Timmermans, M. C. P., Yu, J., Schnable, P. S., & Scanlon, M. J. (2015). Genetic control of morphometric diversity in the maize shoot apical meristem. *Nature Communications*, *6*(1), 8974. https://doi.org/10.1038/ncomms9974

Lelli, V., Belardo, A., & Timperio, A. M. (2021). From Targeted Quantification to Untargeted Metabolomics. In *Metabolomics—Methodology and Applications in Medical Sciences and Life Sciences*. IntechOpen. https://doi.org/10.5772/intechopen.96852

Li, Z., Zhou, P., Coletta, R. D., Zhang, T., Brohammer, A. B., Vaillancourt, B., Lipzen, A., Daum, C., Barry, K., Leon, N. de, Hirsch, C. D., Buell, C. R., Kaeppler, S. M., Springer, N. M., & Hirsch, C. N. (2019). Highly Genotype- and Tissue-specific Single-Parent Expression Drives Dynamic Gene Expression Complementation in Maize Hybrids. In *BioRxiv* (p. 668681). https://doi.org/10.1101/668681

Lin, H., Liu, Q., Li, X., Yang, J., Liu, S., Huang, Y., Scanlon, M. J., Nettleton, D., & Schnable, P. S. (2017). Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. *Genome Biology*, *18*(1), 192. https://doi.org/10.1186/s13059-017-1328-6

Lorenz, A. J., & Coors, J. G. (2008). What can be learned from silage breeding programs? *Applied Biochemistry and Biotechnology*, *148*(1–3), 261–270. https://doi.org/10.1007/s12010-007-8116-9

Lorenz, A. J., Gustafson, T. J., Coors, J. G., & de Leon, N. (2010). Breeding Maize for a Bioeconomy: A Literature Survey Examining Harvest Index and Stover Yield and Their Relationship to Grain Yield. *Crop Science*, *50*(1), 1–12. https://doi.org/10.2135/cropsci2009.02.0086

Lu, F., Karlen, S. D., Regner, M., Kim, H., Ralph, S. A., Sun, R.-C., Kuroda, K., Augustin, M. A., Mawson, R., Sabarez, H., Singh, T., Jimenez-Monteon, G., Zakaria, S., Hill, S., Harris, P. J., Boerjan, W., Wilkerson, C. G., Mansfield, S. D., & Ralph, J. (2015). Naturally p-Hydroxybenzoylated Lignins in Palms. *BioEnergy Research*, *8*(3), 934–952. https://doi.org/10.1007/s12155-015-9583-4

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., Robin Buell, C., de Leon, N., & Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, *19*(1), 1–17. https://doi.org/10.1186/s12870-019-1653-x

Mazur, O., & Zimmer, J. (2011). Apo- and cellopentaose-bound structures of the bacterial cellulose synthase subunit BcsZ. *The Journal of Biological Chemistry*, *286*(20), 17601–17606. https://doi.org/10.1074/jbc.M111.227660

Meghji, M. R., Dudley, J. W., Lambert, R. J., & Sprague, G. F. (1984). Inbreeding Depression, Inbred and Hybrid Grain Yields, and Other Traits of Maize Genotypes Representing

Three Eras [1]. *Crop Science*, *24*(3), 545–549. https://doi.org/10.2135/cropsci1984.0011183X002400030028x

Morgan, J. L. W., Strumillo, J., & Zimmer, J. (2013). Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature*, *493*(7431), 181–186. https://doi.org/10.1038/nature11744

Nachtomy, O., Shavit, A., & Yakhini, Z. (2007). Gene expression and the concept of the phenotype. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *38*(1), 238–254. https://doi.org/10.1016/j.shpsc.2006.12.014

National Research Council. (2001). *Nutrient Requirements of Dairy Cattle: Seventh Revised Edition, 2001*.

Newman, R. H., Hill, S. J., & Harris, P. J. (2013). Wide-angle x-ray scattering and solid-state nuclear magnetic resonance data combined to test models for cellulose microfibrils in mung bean cell walls. *Plant Physiology*, *163*(4), 1558–1567. https://doi.org/10.1104/pp.113.228262

Nixon, B. T., Mansouri, K., Singh, A., Du, J., Davis, J. K., Lee, J.-G., Slabaugh, E., Vandavasi, V. G., O'Neill, H., Roberts, E. M., Roberts, A. W., Yingling, Y. G., & Haigler, C. H. (2016). Comparative Structural and Computational Analysis Supports Eighteen Cellulose Synthases in the Plant Cellulose Synthesis Complex. *Scientific Reports*, *6*(1), 28696. https://doi.org/10.1038/srep28696

Patil, P. S., Fernandes, C. G., Sawant, S. C., Lali, A. M., & Odaneth, A. A. (2022). High-throughput system for carbohydrate analysis of lignocellulosic biomass. *Biomass Conversion and Biorefinery*. https://doi.org/10.1007/s13399-022-02304-8

Pauly, M., Gille, S., Liu, L., Mansoori, N., de Souza, A., Schultink, A., & Xiong, G. (2013). Hemicellulose biosynthesis. *Planta*, *238*(4), 627–642. https://doi.org/10.1007/s00425-013-1921-1

Pauly, M., & Keegstra, K. (2016). Biosynthesis of the Plant Cell Wall Matrix Polysaccharide Xyloglucan. *Annual Review of Plant Biology*, *67*(1), 235–259. https://doi.org/10.1146/annurev-arplant-043015-112222

Pear, J. R., Kawagoe, Y., Schreckengost, W. E., Delmer, D. P., & Stalker, D. M. (1996). Higher plants contain homologs of the bacterial celA genes encoding the catalytic subunit of cellulose synthase. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(22), 12637–12642.

Ralph, J., Brunow, G., Harris, P. J., Dixon, R. A., Schatz, P. F., & Boerjan, W. (2008). Lignification: Are Lignins Biosynthesized via simple Combinatorial Chemistry or via Proteinaceous Control and Template Replication? In F. Daayf & V. Lattanzio (Eds.), *Recent Advances in Polyphenol Research* (pp. 36–66). Wiley-Blackwell. https://doi.org/10.1002/9781444302400.ch2

Ralph, J., Lundquist, K., Brunow, G., Lu, F., Kim, H., Schatz, P. F., Marita, J. M., Hatfield, R. D., Ralph, S. A., Christensen, J. H., & Boerjan, W. (2004). Lignins: Natural polymers from oxidative coupling of 4-hydroxyphenyl- propanoids. *Phytochemistry Reviews*, *3*(1–2), 29–60. https://doi.org/10.1023/B:PHYT.0000047809.65444.a4

Roberts, M. J., Long, S. P., Tieszen, L. L., & Beadle, C. L. (1985). CHAPTER 1—MEASUREMENT OF PLANT BIOMASS AND NET PRIMARY PRODUCTION. In J. Coombs, D. O. Hall, S. P. Long, & J. M. O. Scurlock (Eds.), *Techniques in Bioproductivity and Photosynthesis (Second Edition)* (pp. 1–19). Pergamon. https://doi.org/10.1016/B978-0-08-031999-5.50011-X

Rodríguez-Morató, J., Pozo, Ó. J., & Marcos, J. (2018). Targeting human urinary metabolome by LC-MS/MS: A review. *Bioanalysis*, *10*(7), 489–516. https://doi.org/10.4155/bio-2017-0285

Ruan, Z., Wang, X., Liu, Y., & Liao, W. (2019). Chapter 3—Corn. In Z. Pan, R. Zhang, & S. Zicari (Eds.), *Integrated Processing Technologies for Food and Agricultural By-Products* (pp. 59–72). Academic Press. https://doi.org/10.1016/B978-0-12-814138-0.00003-4

Russel, W. A. (1984). Agronomic performance of maize cultivars representing different eras of breeding. *Maydica*, *29*, 375–390.

Saha, P., Lin, F., Thibivilliers, S., Xiong, Y., Pan, C., & Bartley, L. E. (2021). Phenylpropanoid Biosynthesis Gene Expression Precedes Lignin Accumulation During Shoot Development in Lowland and Upland Switchgrass Genotypes. *Frontiers in Plant Science*, *12*. https://www.frontiersin.org/article/10.3389/fpls.2021.640930

Sakamoto, S., Kamimura, N., Tokue, Y., Nakata, M. T., Yamamoto, M., Hu, S., Masai, E., Mitsuda, N., & Kajita, S. (2020). Identification of enzymatic genes with the potential to reduce biomass recalcitrance through lignin manipulation in Arabidopsis. *Biotechnology for Biofuels*, *13*, 97. https://doi.org/10.1186/s13068-020-01736-6

Santoro, N., Cantu, S. L., Tornqvist, C.-E., Falbel, T. G., Bolivar, J. L., Patterson, S. E., Pauly, M., & Walton, J. D. (2010). A High-Throughput Platform for Screening Milligram Quantities of Plant Biomass for Lignocellulose Digestibility. *BioEnergy Research*, *3*(1), 93–102. https://doi.org/10.1007/s12155-009-9074-6

Schaefer, R. J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., & Myers, C. L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. *The Plant Cell*, *30*(12), 2922–2942. https://doi.org/10.1105/tpc.18.00299

Sekhon, R. S., Lin, H., Childs, K. L., Hansey, C. N., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2011). Genome-wide atlas of transcription during maize development. *The Plant Journal: For Cell and Molecular Biology*, *66*(4), 553–563. https://doi.org/10.1111/j.1365-313X.2011.04527.x

Sethaphong, L., Haigler, C. H., Kubicki, J. D., Zimmer, J., Bonetta, D., DeBolt, S., & Yingling, Y. G. (2013). Tertiary model of a plant cellulose synthase. *Proceedings of the National*

Academy of Sciences of the United States of America, *110*(18), 7512–7517. https://doi.org/10.1073/pnas.1301027110

Shane, M. (2000). The Vascular System of Maize Stems Revisited: Implications for Water Transport and Xylem Safety. *Annals of Botany*, *86*(2), 245–258. https://doi.org/10.1006/anbo.2000.1171

Sharma, A., Shahzad, B., Rehman, A., Bhardwaj, R., Landi, M., & Zheng, B. (2019). Response of Phenylpropanoid Pathway and the Role of Polyphenols in Plants under Abiotic Stress. *Molecules*, *24*(13), 2452. https://doi.org/10.3390/molecules24132452

Soest, P. J. V., & Wine, R. H. (1967). Use of Detergents in the Analysis of Fibrous Feeds. IV. Determination of Plant Cell-Wall Constituents. *Journal of the A.O.A.C.*, *50*(1), 50–55. https://doi.org/10.1016/j.ijhydene.2012.08.110

Somerville, C. (2006). Cellulose Synthesis in Higher Plants. *Annual Review of Cell and Developmental Biology*, *22*(1), 53–78. https://doi.org/10.1146/annurev.cellbio.22.022206.160206

Stanger, T. F., & Lauer, J. G. (2007). Corn Stalk Response to Plant Population and the Bt–European Corn Borer Trait. *Agronomy Journal*, *99*(3), 657–664. https://doi.org/10.2134/agronj2006.0079

Stelpflug, S. C., Sekhon, R. S., Vaillancourt, B., Hirsch, C. N., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2016). An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. *The Plant Genome*, *9*(1). https://doi.org/10.3835/plantgenome2015.04.0025

Strock, C. F., Schneider, H. M., & Lynch, J. P. (2022). Anatomics: High-throughput phenotyping of plant anatomy. *Trends in Plant Science*, *0*(0). https://doi.org/10.1016/j.tplants.2022.02.009

Studer, M. H., DeMartini, J. D., Brethauer, S., McKenzie, H. L., & Wyman, C. E. (2010). Engineering of a high-throughput screening system to identify cellulosic biomass, pretreatments, and enzyme formulations that enhance sugar release. *Biotechnology and Bioengineering*, *105*(2), 231–238. https://doi.org/10.1002/bit.22527

Su, P. (2013). Direct-to-Consumer Genetic Testing: A Comprehensive View. *The Yale Journal of Biology and Medicine*, *86*(3), 359–365.

Tapper, D. C. (1983). *Changes in physiological traits associated with grain yield improvement in single-cross maize hybrids from 1930 to 1970. Ph.D. diss. Iowa State Univ., Ames.* 232.

The Angiosperm Phylogeny Group, Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., Mabberley, D. J., Sennikov, A. N., Soltis, P. S., & Stevens, P. F. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, *181*(1), 1–20. https://doi.org/10.1111/boj.12385

Tilley, J. M. A., & Terry, R. A. (1963). A Two-Stage Technique for the in Vitro Digestion of Forage Crops. *Grass and Forage Science*, *18*(2), 104–111. https://doi.org/10.1111/j.1365-2494.1963.tb00335.x

Tollenaar, M. (1989). Genetic Improvement in Grain Yield of Commercial Maize Hybrids Grown in Ontario from 1959 to 1988. *Crop Science*, *29*(6), 1365–1371. https://doi.org/10.2135/cropsci1989.0011183X002900060007x

Tollenaar, M., McCullough, D. E., & Dwyer, L. M. (1994). Physiological Basis of the Genetic Improvement of Corn. In *Genetic Improvement of Field Crops*. CRC Press.

Upton, B. M., & Kasko, A. M. (2016). Strategies for the Conversion of Lignin to High-Value Polymeric Materials: Review and Perspective. *Chemical Reviews*, *116*(4), 2275–2306. https://doi.org/10.1021/acs.chemrev.5b00345

Valdivia, M., Galan, J. L., Laffarga, J., & Ramos, J. L. (2016). Biofuels 2020: Biorefineries based on lignocellulosic materials. *Microbial Biotechnology*, *9*(5), 585–594. https://doi.org/10.1111/1751-7915.12387

Vermerris, W. (2008). Composition and Biosynthesis of Lignocellulosic Biomass. In W. Vermerris (Ed.), *Genetic Improvement of Bioenergy Crops* (pp. 89–142). Springer New York. https://doi.org/10.1007/978-0-387-70805-8_4

Walley, W. J., Sartor, C. R., Shen, Z., Schmitz, J. R., Wu, J. K., Urich, A. M., Nery, R. J., Smith, G. L., Schnable, C. J., Ecker, R. J., & Briggs, P. S. (2016). Integration of omic networks in a developmental atlas of maize. *Science*, *353*(6301), 0–5. https://doi.org/10.5061/dryad.v8969

Wang, R., Kang, H., Zhang, X., Nie, Q., Wang, H., Wang, C., & Zhou, S. (2022). Urinary metabolomics for discovering metabolic biomarkers of bladder cancer by UPLC-MS. *BMC Cancer*, *22*(1), 214. https://doi.org/10.1186/s12885-022-09318-5

Wehrens, R., Hageman, Jos. A., van Eeuwijk, F., Kooke, R., Flood, P. J., Wijnker, E., Keurentjes, J. J. B., Lommen, A., van Eekelen, H. D. L. M., Hall, R. D., Mumm, R., & de Vos, R. C. H. (2016). Improved batch correction in untargeted MS-based metabolomics. *Metabolomics, 12*, 88. https://doi.org/10.1007/s11306-016-1015-8

Wiersma, D. W., Carter, P. R., Albrecht, K. A., & Coors, J. G. (1993). Kernel Milkline Stage and Corn Forage Yield, Quality, and Dry Matter Content. *Journal of Production Agriculture*, *6*(1), 94–99. https://doi.org/10.2134/jpa1993.0094

Wilson, J. R., & Hatfield, R. D. (1997). Structural and chemical changes of cell wall types during stem development: Consequences for fibre degradation by rumen microflora. *Australian Journal of Agriculture Research*, *48*(2), 165–180. https://doi.org/10.1071/A96051

Wilson, J. R., & Mertens, D. R. (1995). Cell Wall Accessibility and Cell Structure Limitations to Microbial Digestion of Forage. *Crop Science*, *35*(1), 251–259. https://doi.org/10.2135/cropsci1995.0011183x003500010046x

Zhang, Y., Ma, L., Pan, X., Wang, J., Guo, X., & Du, J. (2018). Micron-scale Phenotyping Techniques of Maize Vascular Bundles Based on X-ray Microcomputed Tomography. *Journal of Visualized Experiments: JoVE*, *140*. https://doi.org/10.3791/58501

Zhang, Y., Ma, L., Wang, J., Wang, X., Guo, X., & Du, J. (2020). Phenotyping analysis of maize stem using micro-computed tomography at the elongation and tasseling stages. *Plant Methods*, *16*(1), 2. https://doi.org/10.1186/s13007-019-0549-y

Zhao, Y., Wang, K., Wang, W., Yin, T., Dong, W., & Xu, C. (2019). A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics*, *20*(1), 160. https://doi.org/10.1186/s12864-019-5533-4

Zhou, P., Li, Z., Magnusson, E., Gomez Cano, F., Crisp, P. A., Noshay, J. M., Grotewold, E., Hirsch, C. N., Briggs, S. P., & Springer, N. M. (2020). Meta Gene Regulatory Networks in Maize Highlight Functionally Relevant Regulatory Interactions. *The Plant Cell*, *32*(5), 1377–1396. https://doi.org/10.1105/tpc.20.00080

Zhou, S., Kremling, K. A., Bandillo, N., Richter, A., Zhang, Y. K., Ahern, K. R., Artyukhin, A. B., Hui, J. X., Younkin, G. C., Schroeder, F. C., Buckler, E. S., & Jander, G. (2019). Metabolome-Scale Genome-Wide Association Studies Reveal Chemical Diversity and Genetic Control of Maize Specialized Metabolites. *The Plant Cell*, *31*(5), 937–955. https://doi.org/10.1105/tpc.18.00772

Chapter 2: Genetic dissection and complementarity of stalk anatomy and saccharification efficiency in maize

Authors:

Jonas Rodriguez[1], Marlies Heckwolf[1,2], Shawn M. Kaeppler[1,2,3], Natalia de Leon[1,2]

Affiliations:

[1] Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA

[2] DOE Great Lakes Bioenergy Research Center, 1552 University Ave., Madison, WI 53726, USA

[3] Wisconsin Crop Innovation Center, 8520 University Green, Middleton, WI 53562, USA

2.1 Abstract

Plants, including maize (*Zea mays* L.), are potential feedstocks for lignocellulosic biofuel production. Maize stalks represent the greatest proportion of non-grain biomass but are among the most recalcitrant tissues in physiologically mature maize plants. Recalcitrance is due, in part, to highly lignified anatomical structures such as xylem and phloem. A goal of this study was to assess the relationship between anatomical characteristics in maize stalks and recalcitrance measured as cell wall bound sugar. We also utilized genome-wide association and gene co-expression analysis to identify novel candidate genes associated with anatomical and sugar traits. Genetic analysis was conducted using a set of 532 diverse maize inbred lines and a nested association mapping population consisting of 2,741 recombinant inbred lines from 20 bi-parental families sharing a common parent. We performed enzymatic hydrolysis assays to quantify glucose and pentose from ground tissue as proxies for saccharification efficiency in physiologically mature stalk internodes. Additionally, we assessed ten stalk anatomical features in internode cross-sections of the 532 diverse inbred lines by computational image analysis. Stalk anatomical and sugar traits had moderate to high heritabilities (0.53 – 0.88) among the diverse lines, supporting the potential for genetic dissection. Exemplifying the relationship between anatomical characteristics and recalcitrance, inbreds with smaller stalk areas and densely packed vascular bundles had more glucose. A clustering and connectivity analysis revealed composite phenotypes derived from fermentable sugar traits together with stalk area and vascular bundle density were able to more distinctly separate inbred lines with extreme phenotypes compared to the separation with sugar traits alone. Of the 325 single nucleotide polymorphisms (SNP) significantly associated with stalk saccharification and anatomical traits in both populations, a set of 62 candidate genes were located less than 5kb upstream or downstream from the SNP markers

and were prioritized based on gene co-expression patterns in whole seedling tissue. The functional and pathway annotations for these candidates suggest diverse roles including cell signaling, structure, transport, and transcription factor binding activity.

2.2 Introduction

Approximately 20% of the total annual global energy consumption is directed towards the movement of people and goods by road, air, and water (U.S. Energy Information Administration, 2021). Petroleum-based liquid fuels continue to be the dominant source of energy fueling the transportation sector, though projections through 2050 show an increased proportion of transport energy being derived from both alternative nonliquid fuels such as electric and natural gas, and non-petroleum liquid biofuels and additives such as ethanol (U.S. Energy Information Administration, 2021).

In the United States, efforts to mitigate the reliance on petroleum-based energy sources have primarily been driven by the Energy Independence and Security Act (EISA) of 2007 which, stipulated that almost two thirds of the renewable biofuels requirement should be derived from an advanced biofuel source – a classification defined as non-grain-derived lignocellulosic biomass (cellulose, hemicellulose, lignin), crop residues, and waste materials (U.S. Congress, 2007). Unfortunately, since 2014, EISA advanced biofuel production targets have not been met and the Renewable Fuel Standards (RFS) program has had to adjust production targets and requirements. In addition to low oil prices, leading to unfavorable economics, a major factor contributing to the repeatedly missed production targets include limits on the proportion of ethanol as an additive in fuels (U.S. Energy Information Administration, 2021). For operational lignocellulosic biofuel plants, variability in the biomass supply chain and costs associated with transportation and processing of the source biomass can pose considerable challenges (Golecha

& Gan, 2016). Thus, a reliable biomass source produced close to biofuel plants and with high saccharification efficiency is highly desirable (Kumar et al., 2018).

Multiple plant species have been proposed as bioenergy crop candidates and meeting the overall goal of increasing total lignocellulosic fuel production will require the use of many biomass sources (El Bassam, 2010). Among the top contenders are grass species such as maize (*Zea mays* L.), sorghum (*Sorghum bicolor* L.), and switchgrass (*Panicum virgatum*), which all employ the C4 photosynthetic pathway (Slack & Hatch, 1967). C4 plants are desirable as bioenergy crops due to their high productivity and efficiency in nitrogen use and carbon fixation, which stem from a reduction in photorespiration. In the United States, corn stover is the most abundant and readily available source of lignocellulosic biomass (U.S. Department of Energy, 2011).

On a dry matter basis, corn stover is comprised primarily of cell wall polysaccharides, including the long-chain polysaccharides cellulose (~35%), hemicellulose (~20%) and, the highly recalcitrant plant phenolic polymer, lignin (~12%) (Ruan et al., 2019). Cellulose is a linear homopolymer of glucose (GLU) monosaccharides. In contrast, hemicellulose is a structurally diverse heteropolymer made up primarily of pentose (PEN) monosaccharides linked to various other sugar monomers, including GLU (Brigham, 2018; Scheller & Ulvskov, 2010). When fractionated from cellulose and hemicellulose, GLU and PEN are converted with relative ease to ethanol by microbial fermentation (Fernández-Sandoval et al., 2019; Hoang Nguyen Tran et al., 2020). However, the proportion and composition of lignin in the plant biomass can hinder the saccharification step due to cross-linking with cell wall polysaccharides to form a recalcitrant lignin-carbohydrate matrix complex (Chen & Dixon, 2007; Kang et al., 2019).

Efforts to improve the saccharification step in lignocellulosic biofuel production include using combinations of physical, thermal, chemical, and enzymatic pretreatments (Khare et al., 2015; Kim et al., 2016). Different combinations of pretreatments can have a considerable effect on saccharification. For example, harsh and highly caustic conditions coupled with reducing enzymes tend to yield the most fermentable sugars. However, the inputs necessary to create these conditions are often not economically or environmentally sustainable in practice (Torres et al., 2016). Biomass recalcitrance can be an important factor in the analysis since it is positively correlated with the severity of pretreatment needed. For this reason, generating lignocellulosic feedstocks with improved saccharification under mild pre-treatment conditions is an area of focus in the field of biofeedstock development research (Bichot et al., 2018; Li et al., 2015).

A promising strategy to generate bioenergy crops with desirable saccharification properties is to exploit naturally occurring variation for saccharification related traits, such as GLU and PEN, to aid in the identification of novel candidate genes, which can be subsequently modified by genetic engineering or conventional breeding techniques. Previous research has focused on altering lignin biosynthesis and composition to reduce its negative effect on cell wall sugar accessibility (Barrière et al., 2015). In maize and switchgrass, it has been demonstrated that perturbing genes later in the lignin biosynthetic pathway leads to increased saccharification while maintaining total biomass production (Fornalé et al., 2012; Fu et al., 2011). In comparison, perturbations earlier in the lignin pathway can have deleterious effects on plant growth and development (Pedersen et al., 2005; Yoon et al., 2015). Transcription factors (TFs) play a major role in regulating genes involved in lignin biosynthesis, and might offer opportunities to target various parts of the pathway simultaneously. In addition to TFs, there are likely other still unidentified non-lignin-specific enzymatic genes, that are associated with saccharification traits.

Genome-wide association studies (GWAS) and quantitative trait locus (QTL) mapping for lignification and saccharification traits in maize have been conducted using diversity panels (Penning et al., 2009) and recombinant inbred lines (RILs) (Lorenzana et al., 2010; Penning et al., 2014) from structured populations, respectively. More recently, QTL mapping for saccharification traits using a multi-parent advanced generation intercross (MAGIC) maize population has also been conducted (López-Malvar et al., 2021). A common finding from the studies reporting gene level associations (López-Malvar et al., 2021; Penning et al., 2014) is that TFs are commonly among the top candidate genes associated with saccharification. Interestingly, the Penning et al. (2014) study reported no shared QTL between saccharification and lignification traits, supporting the notion that there may be opportunities for gene discovery and saccharification improvement beyond targeting lignin modifying genes alone.

Differences in tissue anatomy have been speculated to play a major role in cell wall degradability (Wilson & Hatfield, 1997; Wilson & Mertens, 1995). Visualizing anatomical structures in many plant species is tractable using various histological and chemical assays developed to visually and quantitatively assess anatomical structures (Matos et al., 2013; Pradhan Mitra & Loqué, 2014; Tobimatsu et al., 2014). In mature maize plants, stalk internodes can be dissected to reveal distinct anatomical structures and cell types (Shane et al., 2000). The three primary anatomical structures easily visualized are the rind, pith, and individual vascular bundles scattered throughout the rind and pith. Rind tissue consists primarily of small diameter parenchyma cells with thick cell walls and sclerenchyma cells that are more recalcitrant than the larger, thinner-walled cells found in the predominantly parenchymatous pith (Jung et al., 1998). Though histological and chemical assays are effective methods to assess these structures, assaying a large number of plant samples and accurately measuring subtle differences among them can be challenging. To address this challenge, image analysis tools have been developed to increase measurement throughput and resolution (Gomez et al., 2018; Heckwolf et al., 2015; Zhang et al., 2018).

This work investigates the relationships between maize stalk saccharification and anatomical traits, both from a phenotypic and genetic perspective. We explore phenotypic variation and provide insights into leveraging anatomical and biochemical information to characterize phenotypic variation for traits important for bioenergy production. Using high-density molecular markers and genome-wide transcriptional profiling, we conduct GWAS coupled with gene co-expression analysis to dissect the genetic architecture and prioritize candidate genes for relevant traits.

2.3 Materials and Methods

Germplasm and trials

The inbred lines evaluated in this study belong to two maize populations - the Nested Association Mapping population (NAM) (Yu et al., 2008) and the Wisconsin Diversity association panel (WiDiv) (Mazaheri et al., 2019). The NAM population is a set of 25 bi-parental RIL families generated by crossing 25 diverse founder lines with a single common parent, B73. Collectively there are 5,000 recombinant inbred lines (RILs) in the population, with approximately 200 lines from each of the 25 initial F_1's (Yu et al., 2008). The WiDiv population is a set of 942 diverse inbred lines which are adapted to reach grain physiological maturity in the upper Midwest region of the United States. WiDiv lines have previously been categorized into subpopulations by admixture analysis and include the three major heterotic groups of North American field corn, the Stiff Stalk (SS), Non-Stiff Stalk (NSS) and Iodent (IDT), as well as sweet corn, popcorn, and non-temperate tropical lines (Mazaheri et al., 2019).

As part of this study, we evaluated 2,741 RILs from 20 bi-parental NAM families and 532 diverse inbred WiDiv lines (Supplemental Table 2.1). Field experiments were conducted separately for the NAM and WiDiv populations. In all experiments, 20 seeds per maize line were planted in single-row plots 3.80 m long with 0.76 m spacing between plots. The NAM experiment was conducted during the Summer of 2011 at the West Madison Agricultural Research Station (Madison, WI). The field was blocked by bi-parental family and a complete randomization was done within bi-parental family blocks. Each inbred line was represented by one unreplicated plot within its respective bi-parental family block. The WiDiv experiment was conducted during the Summers of 2013 and 2014 at the Arlington Agricultural Research Station (Arlington, WI). These experiments used a randomized complete block design (RCBD) with two replications per line each year. We collected flowering time and saccharification data for both WiDiv and NAM experiments, whereas stalk anatomical traits were collected for the WiDiv experiments alone.

Flowering time and stalk sampling

From each plot, the second and third lowermost stalk internodes were sampled from three separate plants in each plot for phenotyping. The anthesis date was defined by the number of days from planting to when 50% of plants within a plot had started to shed pollen (pollen date), or had silks visibly exposed from ear shoots (silk date). To account for the wide range in anthesis variability, and to control for developmental differences arising from this variability, we collected stalk samples at grain physiological maturity, which we defined as 45 days after the silk date. Our sampling approach involved sorting plots each year, based on silk date, then assigning each plot to one of six sampling groups, where each group represented a silk date window of six days. This was done to distribute the labor associated with sampling while adhering to the targeted developmental sampling time point.

Days to anthesis were converted to cumulative growing degree days (GDD) from planting to pollen-date (GDDP) and silk-date (GDDS) by using the following calculation: $\sum_{i=0}^{n} [(T_{max} - T_{min})/2] - T_{low}$ where n is the days to anthesis, T_{max} is the maximum daily temperature, T_{min} is the daily minimum temperature, and T_{low} is the lower threshold temperature. In this calculation, T_{low} was set to 50°F, and the largest value T_{max} could take was 86°F (Gilmore & Rogers, 1958).

Stalk saccharification traits

Samples were collected from the internode between the second and third above-ground nodes with a high-throughput stalk core sampling device (Muttoni et al., 2012) (Figure 2.1a). Each core was placed in a separate 2 ml screw-cap tube (72.694, Sarstedt, Newton, NC) and stored at -80°C. Once all samples from an experiment had been collected, they were lyophilized for 72 hours prior to shipping to the Michigan State University Cell Wall Facility, where biochemical analysis of glucose (GLU) and pentose (PEN) were performed using a high-throughput digestibility platform (HTDP) (Santoro et al., 2010). The iWALL robotic system is an integral part of the platform and was used for ball mill grinding, weighing, and dispensing samples into three separate sub-samples per 2 ml tube, with a target weight for each sub-sample of 1.5 mg ± 0.2 mg of pulverized biomass (Figure 2.1b). Specific instrumentation and precision metrics of the iWALL system are detailed in (Santoro et al., 2010).

The HTDP platform described in Santoro et al. (2010) was designed to be easily adapted to various pretreatment and reaction conditions. For our materials, the Michigan State University Cell Wall Facility used a mild alkaline pretreatment solution (6.25 mM NaOH) and a two-step biomass hydrolysis incubation. The first incubation was at 90 °C in a water bath for 3 h, and the second incubation was in a rotating hybridization oven (5420, VWR, Radnor, PA) at 50 °C for 20 h. Between the first and second steps, tubes containing the biomass being assayed were cooled on ice, and a 50 µl solution containing a 30 mM citrate buffer plus 0.01% sodium azide and 0.50 µl Accellerase® 1000 (Genencor, Rochester, NY) was added to each tube to neutralize the reaction and to reduce the biomass into fermentable sugars during the second incubation. Following the incubation steps, tubes containing the hydrolyzed biomass were centrifuged at 1,500 x g for 3 min, and the supernatant was assayed in triplicate using the GLU oxidase/peroxidase method (K-GLUC, Megazyme, Ireland) to determine GLU content and photometric measurements using p-bromoaniline as a reagent for PEN determination (Deschatelets & Yu, 1986). GLU and PEN were both normalized to the weight of biomass initially dispensed by the iWALL robotic system for each sub-sample and reported on a percent per milligram dry matter basis (%/mgDM). Thus, for each stalk core sampled, a total of nine technical replicate measurements for GLU and PEN (three per sub-sample) were collected (Figure 2.1b).

Stalk anatomical traits

The same stalk samples from the WiDiv experiments which were cored for the mild alkaline saccharification assays, were used to extract anatomical features by image analysis. Cross sections of the third lowermost internode from each stalk were dissected and imaged on a flat-bed Perfection V700 photo scanner (Epson, Japan). Each scanned image captured 12 internode cross sections representing three separate plants per plot, collected in each of four distinct plots. Each scan was collected at a resolution of 800 dots per inch using the red, green, blue color mode. Images obtained were analyzed using a MATLAB image processing script described in Heckwolf et al. (2015). The image analysis output produced metrics at a plot level for stalk diameter long-axis (SDL) (cm), stalk diameter short-axis (SDS) (cm), average vascular bundle area (VBA) (cm2), vascular bundle density (VBD) (bundles/cm2), and rind thickness (RT) (cm) (Figure 2.1a). Using these measurements, and assuming the shape of an ellipse, we calculated the following derived traits: stalk area (SA) (cm) = π (1/2 SDL* 1/2 SDS), pith area (PA) (cm^2) = π (1/2 (SDL-RT)* 1/2 (SDS-RT)), total number of pith bundles (TB) (counts) = PA*VBD, total pith bundle area (TBA) (cm^2) = TB * VBA, and rind area (RA) (cm^2) = SA-PA.

Phenotypic analysis

Phenotypes were analyzed separately for the WiDiv and NAM populations due to differences in the experimental designs. In both cases, however, the experimental unit was the trait value on a per-plot basis. Flowering time was recorded as a single value per plot for pollen-date and silk-date, and in turn, flowering time phenotypes GDDP and GDDS consisted of single measurements per plot. For biochemical saccharification assays, the nine technical replicate measurements from each of the three stalk cores were individually averaged, resulting in three plant level measurements per plot. A pairwise difference outlier detection method was applied to these measurements prior to averaging within a plot to produce a plot-level value.

The outlier detection method was performed according to the following steps: first, the pairwise absolute differences were calculated between plant samples within each plot, then these differences were compared among pairwise differences in all plots evaluated for that year. If a given difference was greater than the 0.95 percentile of differences among plots, then the individual plant measurement with the most extreme value compared to the other two from the same plot was eliminated.

Individual plant measurements for anatomical traits, measured on the WiDiv alone, were averaged on a per-plot basis by the image processing pipeline. Therefore, no individual plant level outlier detection was conducted on this data. The set of anatomical traits returned by the image analysis tool (Heckwolf et al., 2015) have previously been published on the WiDiv in Mazaheri et al. (2019). However, the data published there was reported as trait BLUPs and stalk diameter traits were combinations of manual and image based measures. In this study we used plot level measures for many analyses, therefore for consistency we re-analyzed and provided data as plot level values for all traits (Supplemental Table 2.1).

An additional data cleaning step was applied to the phenotypic data collected on the WiDiv population, whereby inbred lines were completely excluded from the analysis for all traits if data were not available for at least two plots between years and replications. This filtering was done to reflect the multi-year replicated design of the experiment.

All descriptive statistics, including trait distributions, comparisons between groups, and Pearson correlations (ρ) among traits used plot-level trait measurements. We conducted tests of normality using the Shapiro-Wilk test (W) with a type 1 error threshold of 5% and assessed significance for correlations and between group comparisons with a two-sided T-test (T) using the same error threshold as W.

Phenotypic variances were decomposed by analyzing the WiDiv population alone, since the NAM population was evaluated using an unreplicated design where variance partitioning is not possible. The following linear mixed model (LMM) was used to analyze the WiDiv phenotypic data: $Y_{ijk}= \mu + E_i + G_j + R_{k(i)} + (GE)_{ij} + \varepsilon_{ijk}$ where Y_{ijk} is the plot-level response variable in the ith year of the jth genotype in the kth replication. μ is the grand mean, E_i is the ith year, G_j is the jth genotype, $R_{k(i)}$ is the kth replication nested in the ith year, $(GE)_{ij}$ is the interaction between the j^{th} genotype in the i^{th} year and ε_{ijk} is the residual error. All factors were treated as random effects except for μ, which was treated as a fixed effect. Variance components were estimated by restricted maximum likelihood (REML) and subsequently used to calculate the entry mean narrow sense heritability of each trait as $h^2 = \sigma_G^2 / \sigma_G^2 + (\sigma_{GxE}^2/e) + (\sigma\varepsilon^2/re)$ where σ_G^2 is the genotypic variance, σ_{GxE}^2 is the genotype-by-year interaction variance, $\sigma\varepsilon^2$ is the error variance, e is the number of years and r is the number of replications per year. The significance of each model term was tested using a Chi-square distribution with the REML likelihood ratio (LR) as the test statistic and a test type 1 error threshold of 5%. If σ_{GxE}^2 was not significant for a given trait, best linear unbiased predictors (BLUPs) were calculated from the mixed model. When σ_{GxE}^2 was significant, the Spearman rank correlation between years for the trait in question was calculated, and if the correlation was significant (Pr>T <0.001), then we proceeded with the BLUP calculation. Mixed models were fitted using the lme4 R package (Bates et al., 2015), and significance testing was done with the lmerTest R package (Kuznetsova et al., 2017).

To underscore the interplay between stalk anatomical and saccharification traits, we reduced the dimensionality of the phenotypic data collected on the WiDiv population. Using centered and scaled trait BLUPs (mean = 0, variance = 1), we performed a principal components analysis (PCA) and retained the first two principal components (PCs) to represent the observed phenotypic variation (phenotypic space). In addition to the 12 independent saccharification and anatomical phenotypes used to calculate the first two PCs, we derived two composite phenotypes by combining GLU with SA and VBD, designated as [GLU, SA, VBD], and by combining the same two anatomical traits with PEN, [PEN, VBD, SA]. The trait combination was done per inbred line by taking the average percentile rank of each set of trait BLUPs within the brackets shown above. Our intention with these two composite phenotypes was to capture variation related to saccharification efficiency while simultaneously considering the importance of stalk size and vasculature.

We developed and implemented a clustering via connectivity analysis on the phenotypic space, with an emphasis on inbred lines at the tails (top and bottom 25 lines) of individual or composite phenotype distributions. First, we visually assessed the size and coordinate positions of the convex hulls formed by each phenotypic extreme group of inbreds The convex hull size represented the clustering strength, where smaller hulls suggested tighter clustering. Spatially, greater Euclidian distances between all pairwise between group points (inbreds) were interpreted as a measure of distinctiveness between groups.

Secondly, for a subset of phenotypes (GLU, PEN, SA, VBD, [GLU, SA, VBD], [PEN, VBD, SA]), we took a dynamic and quantitative approach to determine the persistence of within to between group clustering and connectivity, by relying on principles from topological data analysis (Chazal & Michel, 2021; Edelsbrunner et al., 2000). For each phenotype, each inbred line belonging to one of two phenotypic extreme groups was represented in the phenotypic space by a 0-simplex (a point with a diameter of 0.50). The diameter of all points was increased incrementally from 0.50 to 16 by a unit of 0.10, and the total number of 1-simplices (point-to-point connections) formed by an overlap of any two points was recorded at each point diameter increment. We used the frequency of within to between group point-to-point connections as a function of point diameter size to compare clustering and connection patterns between the traits analyzed. Visualization and animation were done using the ggtda R package and ImageMagick software (v.6.9.7-4), respectively (Brunson et al., 2021; The ImageMagick Development Team, 2021).

Genotypic and gene expression data

Marker data for the WiDiv was obtained from (Mazaheri et al., 2019). The file consists of 899,784 SNPs called from aligning RNA-seq reads to the B73 AGPv4 reference genome. A minor allele frequency (MAF) filtering was done to retain SNP markers with MAF ≥ 0.05, which resulted in a SNP matrix of 426,149 x 532 inbred lines. The genotypic data for NAM lines was downloaded from the CyVerse Data Store (/iplant/home/shared/panzea/VCAP/genotypes/NAM/namrils_projected_hmp31_MAF02mnCnt2500.hmp.txt.gz). These SNP markers were called using the B73 AGPv3 reference genome. Therefore for comparable coordinates to WiDiv SNPs, markers were uplifted to AGPv4 coordinates using the Picard tools (v.2.23.8) LiftoverVcf utility (Broad Institute, 2019), and the AGPv3 to AGPv4 assembly chain file (Ensembl Plants v48, 2019). After applying the same MAF filtering criteria as applied to the WiDiv data, the resulting SNP matrix consisted of 900,260 SNP markers x 2,741 RILs.

Gene expression data (V1 developmental stage whole seedling tissue) for the full WiDiv population (942 lines) was obtained from (Zhou & Springer, 2020) as raw counts per million (CPM) aligned to the B73 AGPv4 reference genome (Jiao *et al.* 2017). The gene expression data repository published by Zhou and Springer (2020) were produced by re-analyzing the same RNA-seq reads published and used in Mazaheri et al. (2019) to generate the WiDiv SNP data referenced above. We further normalized the CPM data by the annotated (Ensembl Plants v48, 2019) gene lengths to give gene expression values as transcripts per million (TPM).

Population structure and kinship analyses

We assessed population structure (Q) and kinship (K) separately in the WiDiv and NAM populations using their respective SNP matrices. For analyses regarding Q, we first filtered SNPs by binning them into 200 kb bins followed by an iterative SNP removal from each bin if the maximum squared correlation between a given SNP and another SNP in the same bin was ≥ 0.5. The filtered SNP set represented linkage pruned (LD) SNPs since within the 200 kb windows, only the least correlated (linked) were retained. We used LD pruned SNPs only for assessing Q, not for association mapping, since with LD pruning there is a possibility of removing true genetic signals. The LD pruning was done with the snp_autoSVD function from the bigsnpr R package (Prive et al., 2018). For the WiDiv, 118,135 SNPs remained after LD pruning, and for the NAM, this number was 80,698. Q was represented by the first 5 PCs generated from PCA on LD-pruned SNPs, while the pairwise covariance matrix of MAF filtered SNPs was computed to represent K.

The characteristics of Q and K in both WiDiv and NAM populations has been reported in the literature (Mazaheri et al., 2019; Yu et al., 2008). The goal with this analysis was to compute SNP derived PCs for both populations using an independent SNP sampling strategy and to use them as covariates in GWAS. For the WiDiv, the first 5 PCs explained 14.8% of the variation and clearly separated 12 previously reported sub-populations (Supplemental Figure 2.1a). Similarly, for the NAM the top 5 SNP derived PCs explained 16.3% of the variation and provided adequate separation among the RIL families (Supplemental Figure 2.1b).

Genome-wide association studies

Genome wide association studies (GWAS) were performed using trait BLUPs for the WiDiv and plot-level measurements for the NAM. We conducted seven separate GWAS analyses for each trait measured in each population, with varying levels of K and Q accounted for as covariates. A single marker analysis was used to construct our null model without any control for K or Q, and for the remaining six GWAS models fit, K was included in each, while Q was varied from including 0PCs to including up to 5PCs. These seven models were designated as Naive, K, K+Q (PC1), K+Q (PC1:PC2), K+Q (PC1:PC3), K+Q (PC1:PC4), K+Q (PC1:PC5). Models which included Q and K were fitted using the iterative FarmCPU method for both the WiDiv and NAM (X. Liu et al., 2016).

Briefly, FarmCPU develops a multi-locus mixed model by iteratively fitting fixed and random effects models separately. This is done by first fitting the fixed effects model with Q independently, then the genome is split into bins, and pseudo trait associated SNPs (TASs) in each bin are identified by single marker analysis. The pseudo TASs are then used to define K among lines, and this K is used as a covariate in the fixed effects model of the next iteration. The process is repeated until no new pseudo TASs are identified. FarmCPU has previously been used to dissect other phenotypes in these two populations and has been shown to increase statistical power relative to other GWAS methods by reducing false positives while simultaneously avoiding false negatives (Kusmec et al., 2017; X. Liu et al., 2016; Mazaheri et al., 2019). We used all default FarmCPU parameters, with the exception of bin sizes for bin optimization which was set to 5, 10, 50, and 100kb, and the number of iterations run was increased from the default of 10 and set to 20. The implementation of FarmCPU was done using the rMVP R package (Yin et al., 2021).

Identifying and prioritizing candidate genes

For each phenotype, we developed a simple aggregation method (Mult-Mod-Avg) to summarize GWAS results across all seven models fit. This was achieved with a four-step filtering and querying process (Supplemental Figure 2.2). First, TASs were retained if they surpassed a liberal p-value threshold of 1/M, where M is the number of markers tested in any of the GWAS models fitted. Secondly, we filtered for TASs which persisted in at least two models.

Thirdly, an average p-value was calculated after removing naive model results since these tend to have hyper-inflated p-values. And lastly, a maximum distance of 50kb up/downstream between the SNP position and a given gene's transcription start site (TSS) was used to identify candidate genes.

We applied this approach using GDDS data collected on the WiDiv, with the presumption that previously identified flowering time genes would be clearly identified. To statistically test the approach, we used the WiDiv SNP data to simulate nine phenotypes with differing proportions of phenotypic variance explained by the same 50 causal SNPs (Vsnp), genetic background effects (Vbg) (e.g., Q and K), and error (Verr). Simulations were done using the R package PhenotypeSimulator (Meyer & Birney, 2018). Here the sum of Vsnp and Vbg represented a simulated measure of h^2. The specific values of Vsnp, Vbg, and Verr for the nine phenotypes were designated with an underscore symbol between them, and included: a high h^2 group (0.9_0_0.1), (0.6_0.3_0.1), (0.3_0.6_0.1), a moderate h^2 group (0.6, 0, 0.4), (0.4, 0.2, 0.4), (0.2, 0.4, 0.4), and a low h^2 group (0.3, 0, 0.7), (0.2, 0.1, 0.7), (0.1, 0.2, 0.7). Each phenotype was simulated five times with a new seed argument, and GWAS was conducted for each simulation as outlined in the previous section above. The efficacy of our Mult-Mod-Avg method to control for type 1 and type 2 errors was compared to running each of the GWAS models independently. To compare between models, we used the least-square means (Lenth, 2016) of the area under the smoothed receiver operator characteristic (ROC) curve for each model. ROC curves were constructed using the pROC R package (Robin et al., 2011), with $-\log_{10}$ (p-values) from GWAS model fits as predictors.

We applied our Mult-Mod-Avg method to our real anatomical and saccharification GWAS results. Candidate genes identified from TASs within a small distance (5kb) and surpassing a conservative Bonferroni correction significance threshold (-$\log_{10}$ (p-value) = 6.93) used for both WiDiv and NAM) were investigated for co-expression patterns using TPM gene expression data from the WiDiv. Co-expression affinity was assessed using genome-wide mutual rank Pearson correlation coefficients (MRC) between gene pairs. MRCs are relative values between gene pairs that take into consideration genome-wide co-expression patterns. For instance, if we consider the expression of two genes (gene A and gene B), Pearson correlations are calculated between gene A and all genes genome wide, then the same is done for gene B. The MRC is then calculated as the geometric mean between the Pearson correlation rank of gene A with gene B and gene B with gene A (Obayashi & Kinoshita, 2009).

In our analysis, we used a set of 34,881 genes as the genome wide comparison, produced by taking the 46,023 reference gene models (Ensembl Plants v48, 2019) and filtering out genes with zero variance in expression and those where more than 90% of lines had a TPM value of zero. MRCs between gene pairs were scaled to have values between 0 and 1, and the top 5% for each gene were retained, and a gene-clique clustering was done based on edge-betweenness centrality scores (Girvan & Newman, 2002). Accordingly, we designated high priority candidate genes FOR THIS STUDY as those which were identified by GWAS and were also highly co-expressed with other GWAS identified genes after taking into account background genome-wide co-expression patterns GIVEN THE FOCUS ON PHENOTYPIC EXTREMES. For high priority candidates with an unknown function in maize, we queried the PaperBLAST database for orthologs in different plant species (Price & Arkin, 2017). From our set of refined candidate genes, we focused our discussion on genes with functional characterization, either in maize or other species, which related directly to biomass composition, growth and development, or the transcriptional regulation of genes governing these processes. For candidate genes identified as TFs, we included both the reference gene model name (Ensembl Plants v48, 2019) and Grassius nomenclature (Gray et al., 2009; Yilmaz et al., 2009).

2.4 Results

Trait means and distributions

Among traits measured in both the WiDiv and NAM, we observed significant differences when comparing between populations for GLU, GDDP, and GDDS (Figure 2.2). On average, NAM lines were more recalcitrant (1.4 %/mg less GLU) and flowered later (+376 GDDP and +414 GDDS), compared to WiDiv lines (Figure 2.2). We observed significant between-year differences among WiDiv lines alone for GLU, GDDP, and GDDS, with GLU in 2014 being on average 1.06 %/mg higher than in 2013, and flowering time in 2014 occurring later (+62 GDDP and +58 GDDS) than in 2013 (Figure 2.2). Raw data distributions of flowering time (both GDDP and GDDS) significantly deviated from normality (Pr<W<0.05) in both WiDiv and NAM lines. Though, the deviations from normality were more pronounced in the NAM where the distribution patterns for GDDP and GDDS appeared to be multimodal (Figure 2.2). There were no significant differences between years for any of the stalk anatomical traits evaluated in the WiDiv. However, between replicate differences were significant for SDL and SA in 2013 and 2014, and for SDS, PA, and TB in 2014 alone (Figure 2.2).

Phenotypic correlations

The majority of correlations between all trait pairs within saccharification and anatomical traits independently, then between saccharification and anatomical traits, saccharification and flowering time, and flowering time and anatomical traits were statistically significant (Pr > T < 0.05). The non-significant correlations included GLU with TBA, PEN with TB, and VBA with both GDDP and GDDS.

A strong positive correlation was observed between GLU and PEN in both the WiDiv and NAM populations (Figure 2.3). When analyzed across populations, the correlation between GLU and PEN was ρ = 0.70, which did not considerably differ from when the data were analyzed separately by population, ρ = 0.72 for the NAM and ρ = 0.73 for the WiDiv (Figure 2.3). The linear relationship between GLU and PEN was consistent with our expectations and suggests the mild alkaline hydrolysis assay used to extract both sugars did so at similar rates across all ground stalk internode samples assayed.

Collinearity among the stalk anatomical traits evaluated in the WiDiv was persistent, with correlations as high as ρ = 0.97, for example, between SA and both SDS and SDL (Supplemental Figure 2.3). The high proportion of correlated anatomical traits was not unexpected given that half were derived using SDL, SDS, RT, VBA, and VBD values. The derived traits served to both reinforce and further dissect relationships between traits directly measured. For instance, correlations between VBD and traits related to stalk size (SDL, SDS, SA) were all strongly negative (ρ = -0.65, ρ = -0.67, ρ = -0.67 respectively), suggesting larger stalks have a greater dispersion of vascular bundles. Furthermore, since VBD was measured specifically for vascular bundles in the stalk pith, the negative correlation between VBD and PA (ρ = -0.45) supports the relationship between VBD and stalk size. However, bundle dispersion and stalk size characteristics alone leave out a crucial aspect of the vasculature in stalk internodes, which is the counts of TB present. The correlations between TB and the same set of traits related to stalk size were all moderately positive, ρ = 0.24 with SDL, ρ = 0.22 with SDS, and ρ = 0.22 with SA. Taken together, the VBD and TB correlations with stalk size traits indicate that although VBD is lower in larger stalks, there is typically also a greater number of TB in large stalks compared to small stalks. This relationship is particularly relevant considering vascular bundles in physiologically mature maize plants are considered to be highly recalcitrant due to the partial or complete lignification of the cell walls which make up the bundles.

Generally, correlations between GLU and PEN with stalk anatomical traits were weak to moderate and negative. The exceptions to this were VBD, which was positively correlated with both GLU (ρ = 0.22) and with PEN (ρ = 0.19), and TB, which was only weakly correlated with GLU (ρ = 0.08) (Supplemental Figure 2.3). The positive correlations between both saccharification traits and VBD were unexpected, and imply that a mild alkaline pretreatment of the stalk biomass may be sufficient to eliminate the expected detrimental effects associated with a high VBD on saccharification. Additionally, the lack of a significant correlation between TB and PEN and a low correlation with GLU supports this notion.

Flowering time traits are among the most ubiquitously collected in maize evaluation experiments. In each of our experiments, we used flowering time as a proxy for controlling developmental differences among lines when sampling stalks. Therefore, our expectation was that by controlling for flowering time, we would have mitigated the confounding effects of GDDP and GDDS on any of the stalk traits evaluated. Controlling for this confounder was particularly imperative for the saccharification assays given the dynamic nature of sugar biosynthesis, translocation, and accumulation through development.

The relationship between GLU and PEN with GDDP and GDDS was inconsistent between the NAM and WiDiv (Figure 2.3). In the WiDiv, GLU was negatively correlated with GDDP and GDDS (ρ = -0.44 for both), and PEN was also negatively correlated with both (ρ = -0.40 and ρ = -0.39, respectively). Conversely, in the NAM, correlations between GDDP and GDDS with GLU were near zero and non-significant, and with PEN very weak (ρ = -0.03 and ρ = -0.05, respectively). When the data from both populations were analyzed together, the correlations between GLU and both GDDP and GDDS remained moderately negative (ρ = -0.36 for both) while for PEN, correlations were fairly weak (ρ = -0.09 with GDDP and ρ = -0.08 with GDDS). Thus, suggesting that while a relatively small number of earlier flowering lines from the WiDiv seem to be driving the negative correlations between saccharification and flowering time, this occurs to a greater extent with GLU compared to PEN.

Correlations between flowering time and anatomical traits in the WiDiv were weak to moderate and generally positive, which contrasts with the predominantly negative correlations between anatomical and saccharification traits (Supplemental Figure 2.3). The weakest correlations between GLU and PEN with anatomical traits (TB, TBA) were different from the weakest correlations between flowering time and anatomical traits (VBD, VBA) (Supplemental Figure 2.3). This finding may indicate that the biological processes governing stalk anatomical traits such as VBD and VBA, which are important for GLU and PEN, are not necessarily related to the same processes governing flowering time. Or that some traits are more prone to technical errors and exhibit more non-genetic variation which results in reduced correlations.

Variance decomposition

The multi-year replicated experimental design of the WiDiv allowed for variance decomposition of traits and subsequent estimation of entry mean heritabilities. There was substantial genetic variation among lines and moderate to high estimates of entry mean heritabilities for all traits, ranging from 0.58 for RT to 0.97 for GDDP and GDDS (Table 2.1). Genotype-by-year effects were significant (Pr>LR <0.05) only for GLU and PEN and GDDP (Table 2.1), however, Spearman rank correlations between years for these traits were also significant (Pr>T <0.001) and positively correlated, 0.55 and 0.44 for GLU and PEN, respectively, and 0.91 for GDDP (Supplemental Figure 2.4). Therefore, for uniformity in downstream analyses we calculated BLUPs for all traits using the same LMM to analyze the data across years and replications. Our assumption was that even for traits with significant genotype-by-year effects, inbred lines at the extremes of the distributions would remain consistent, and therefore not introduce spurious results from analyses that are driven by extreme observations.

Phenotypic extremes

The value of having both anatomical and saccharification phenotypes for the WiDiv was most apparent when reducing the dimensionality of our dataset and focusing on inbred lines at the tails of each phenotypic BLUP distribution. Together, the first and second PCs derived from phenotypes explained 64.9% of the variability in the data, with PC1 explaining 44.3%, and PC2 explaining 20.6% (Supplemental Figure 2.5).

For GLU, PEN, SA, VBD, and the two composite phenotypes, [GLU, SA, VBD] and [PEN, VBD, SA], we examined static and dynamic clustering patterns of individuals expressing extreme phenotypes. Based on the size of the convex hull alone, the composite phenotypes appeared to cluster most tightly within the phenotypic extreme groups, and they occupied a space distinct from either of the stalk anatomical and saccharification phenotypes alone (Figure 2.4a). SA had the largest separation between the top and bottom 25 lines, followed closely by VBD (Figure 2.4a). A small but noticeable proportion of lines with extreme GLU and PEN phenotypes overlapped in the PC space, suggesting that anatomical traits are likely contributing a greater proportion to the phenotypic separation among lines (Figure 2.4a).

The dynamic clustering approach of using within to between group point connections as a function of increasing individual point diameters in the PC space allowed for a quantitative assessment of within group clustering persistence (Supplemental File 2.1). A within group frequency of 1 indicated a complete separation of the top and bottom 25 lines into two distinct groups, while a frequency of 0.5 marked the point at which the top and bottom 25 lines are all connected, and a single indistinguishable group was formed. For SA and VBD, a within group frequency of 1 was maintained for point diameter sizes up to 6.75, and 1.55, respectively (Figure 2.4b, Supplemental File 2.1). In contrast, GLU and PEN within group frequencies over 0.94 were never reached, and the maximal frequency was achieved when point diameters were small (<0.90), and few total connections had been made (Supplemental File 2.1). Within group frequencies of 1 were maintained for both composite traits for point diameter sizes up to 1.3 (Figure 2.4b).

The relatively low within group connectivity frequency observed for saccharification traits was not unexpected, given the overlapping convex hulls of the two extreme groups for both GLU and PEN (Figure 2.4a). The rate at which the within group frequency declined was greatest for SA, which interestingly can also be interpreted as a proxy for group clustering strength, given that a more rapid decline occurs when the starting points are closer to one another. Rates of decline were lowest for GLU and PEN, and intermediate for the two composite traits (Figure 2.4b). Connectivity frequencies and their rate of decline might be indicative of the underlying biological processes that lead to the observed phenotypes. For instance, comparing the fast rate of decline for SA to the slow rate of PEN might suggest that PEN is less deterministic than SA and possibly more genetically complex to dissect.

Candidate genes for traits relevant to bioenergy

We tested our Mult-Mod-Avg GWAS result summarization approach by applying it to GDDS data collected from the WiDiv population. As expected, this method was able to identify two previously cloned flowering time genes (Liang et al., 2019), one of which corresponded to the most significant association in the test analysis (Supplemental Figure 2.2). When applying our method to simulated data, we found that it performed as well, or marginally better than any single GWAS model (Supplemental Figure 2.6). Importantly though, it did not perform worse, and it provided a non-subjective approach to correcting for Q and K.

Following the same approach and applying it to GWAS conducted on the NAM and WiDiv, we identified a total of 864 candidate genes corresponding to 325 unique TASs associated with at least one stalk saccharification or anatomical phenotype (Supplemental Table 2.2, Supplemental Figure 2.7). Given the relatively large number of total candidate genes identified, it would be unreasonable to assume each as a potential candidate warranting further investigation. Therefore, to prioritize candidate genes and to emphasize cases where there is complementarity between anatomical and saccharification traits, we focused on a set of 50 TASs associated with either GLU and PEN (in both WiDiv and NAM), SA, VBD, [GLU, SA, VBD] and [PEN, VBD, SA] (WiDiv alone). This set of TASs surpassed a conservative Bonferroni correction significance threshold ($-\log_{10}$ (p-value) = 6.93) and were individually less than 5kb from the TSS of a protein coding gene. An additional criterion for the high priority candidate genes was for them to be within the top 5% of MRCs based on genome-wide co-expression. Thus, with this refined criteria, 62 high priority candidate genes corresponding to 18 gene cliques remained (Figure 2.5).

Candidate genes for GLU and PEN accounted for ~70% of all high priority candidates identified, with 28 and 17 genes identified, respectively. These candidates were separated into 9 gene cliques, 4 for GLU and 5 for PEN (Figure 2.5a). Within the 4 GLU gene cliques, WiDiv and NAM identified candidate genes appeared in each, whereas for PEN, there were 3 population specific cliques and the majority of candidate genes were identified in the NAM population (Figure 2.5a). Between GLU and PEN, two candidates were shared, these included

Zm00001d031443 and Zm00001d031444, which were both identified in the NAM population alone. The gene function of both Zm00001d031443 and Zm00001d031444 remains uncharacterized in maize, though an ortholog of Zm00001d031443 in Arabidopsis (At5G43720) has been found to tightly cluster with GIGANTUS1 (GST1), a transducin protein which has been shown to negatively affect biomass yield (Gachomo et al., 2014). For candidate genes not overlapping between GLU and PEN, there was extensive interconnectedness between gene nodes, suggesting gene regulatory processes at play. This was supported by 4 GLU and 1 PEN candidate genes annotated as encoding TFs: ZmEREB198 (Zm00001d002762), ZmHB41 (Zm00001d017422), ZmCAMTA3 (Zm00001d042313), ZmDOF19 (Zm00001d042736), and ZmC3H30 (Zm00001d034710), respectively.

Gene cliques formed by high priority candidate genes for anatomical and composite traits had fewer candidate gene members compared to both saccharification traits, with at most 3 candidate genes making up a single gene clique (Figure 2.5b & Figure 2.5c). There were a greater number of high priority candidate genes identified for SA (7 genes) and VBD (5 genes) compared to [GLU, SA, VBD] (4 genes) and [PEN, VBD, SA] (3 genes). However, all genes for the composite traits were connected based on co-expression by at least one edge, opposed to SA and VBD, which each had isolated gene cliques (Figure 2.5b & Figure 2.5c).

The few candidates identified for [GLU, SA, VBD] and [PEN, VBD, SA] are particularly relevant given that the composite traits might represent a specific combination of desirable traits for the improvement of bioenergy crops. The high priority candidates identified for [GLU, SA, VBD] included the TF ZmARF20 (Zm00001d015243), a bZIP TF-like protein (Zm00001d038223), and two uncharacterized genes (Zm00001d015247, Zm00001d015248). The ZmARF20 (Zm00001d015243) ortholog in Arabidopsis (At1G19220) has been demonstrated to serve as a transcriptional activator when a cofactor (At5G20730) is present (Wilmoth et al., 2005). Functional characterization has recently demonstrated that in actively growing cells, the two ARF proteins produced from At1G19220 and At5G20730 are localized to the nucleus, where they modulate auxin responsiveness. However, in cells not actively elongating, the regulatory proteins are partitioned to the cytoplasm, where they have a reduced effect on the transcriptional landscape of their respective target genes (Powers et al., 2019). In our study ZmARF20 (Zm00001d015243) was positively correlated and mutually co-expressed with two other [GLU, SA, VBD] candidates, Zm00001d038223 ($\rho = 0.30$, MRC = 0.85) and Zm00001d015248 ($\rho = 0.53$, MRC = 0.83) (Figure 2.5c). Additionally, the SNP variation within the gene body of ZmARF20 (Zm00001d015243), which allowed us to identify this gene through GWAS, may serve as an immediately available source of natural variation to exploit.

For [PEN, VBD, SA], the high priority candidates identified were all non-TF genes and included Zm00001d029527, Zm00001d001850, and Zm00001d015242. Among these, the most relevant to our traits of interest was Zm00001d029527, which encodes a small (107 amino acid) SPIRAL1-like protein. This same protein has been previously shown in Arabidopsis (At1G26355) to play an important role in cortical microtubule orientation which is a crucial aspect of cellulose microfibril deposition in cell walls during growth and development (Baskin, 2001; Nakajima et al., 2004, 2006). Similar to the top candidate for [GLU, SA, VBD], the TAS leading us to Zm00001d029527 was found within the gene body, only 226 bp from its TSS.

2.5 Discussion

Stalk anatomy and biochemistry are complementary in the context of bioenergy crop improvement

Biochemical assays such as GLU and PEN are undoubtedly important for screening and identifying high quality sources of fermentable biomass to increase saccharification efficiency (Khare et al., 2015). Additional factors such as tissue anatomy can also be important for determining biomass degradability and that impact would not be fully reflected in biochemical assays alone (Barros-Rios et al., 2012). Furthermore, saccharification can be dependent on extrinsic factors such as pretreatments and thermal conditions, which may situationally alter the definition of what constitutes an ideal biomass source (Galbe & Wallberg, 2019). For instance, under mild thermochemical conditions, a positive correlation was found between fermentable sugar and VBD (Figure 2.3a). This was unexpected, given the negative relationship between cell wall degradation and the composition of cell types making up the internode vasculature (Jung et al., 1998) and that this relationship is most notable in tissues that have reached physiological maturity (Morrison et al., 1998). With milder conditions or pretreatments compared to those used in our study, it is likely that the correlation between VBD and saccharification would be diminished or potentially reversed. Yet given the positive relationship observed under our evaluation conditions, we observed that composite phenotypes are able to separate phenotypically extreme individuals nearly as well, or in some cases better than a single phenotype (Figure 2.4).

In some circumstances, individuals with large stalks and high saccharification might be more desirable than those with narrow stalks and low saccharification. In maize production for grain, breeding for and planting at higher densities has contributed to substantial grain yield increases over time (Duvick, 2005). However, at high densities, a crop is more vulnerable to stalk lodging, which can diminish gains from utilizing a high planting density (Duvick et al., 2003; Stanger & Lauer, 2007). Under these conditions, stover yield can be negatively affected, which might be a contributing factor to the vulnerability to stalk lodging. When used as a dual-use crop (grain and stover for biorefining) high yields of grain and fermentable stover are desirable (Barrière et al., 2015). Thus, varieties developed from individuals with high saccharification efficiency and large stalks that are potentially more tolerant to stalk lodging might be optimal for certain management practices and production goals.

Reducing the intrinsically multidimensional nature of phenotypes

Dimension reduction by PCA has been widely used in biology as an exploratory and grouping analysis technique (Ringnér, 2008). Motivated by the high degree of phenotypic collinearity observed among the traits evaluated in our study, particularly anatomical traits (Supplemental Figure 2.3), we applied PCA to reduce and represent stalk phenotypic variability with the first and second PCs. PCA with biomass quality related phenotypes has previously been used to differentiate germplasm origin (Munaiz et al., 2021). However, PCA clustering alone does not give a quantitative assessment of the strength of clusters that may visually appear. Combining PCA with low dimensional topological data analysis of phenotypic extremes for GLU, PEN, SA, VBD, [GLU, SA, VBD], and [PEN, VBD, SA] we found that composite phenotypes separated extreme individuals more distinctly than saccharification phenotypes but less than anatomical phenotypes (Figure 2.4, Supplemental File 2.1). The application of topological data analysis to investigate biological questions is still in its infancy (Amézquita et al., 2020), and in our study, we investigated only single dimensional features (point-to-point connections) as a means to quantitatively assess inter and intragroup distinctness (Figure 2.4, Supplemental File 2.1). We also hypothesized that there might be an inverse relationship between trait complexity and the within group connectivity frequency decay rate (Figure 2.4b, Supplemental File 2.1). When basing our assessment of biological complexity on the number of high priority genes identified, the relationship between the two was not clear. In the WiDiv, PEN had the slowest decay rate (Figure 2.4b), but we identified fewer high priority candidate genes for PEN than we did for SA, which had the fastest decay rate (Figure 2.5). Also in the WiDiv, GLU had a relatively low rate of decay, but we identified the most number of candidates (Figure 2.5). A plausible explanation might be that the number of high priority candidate genes identified does not necessarily reflect the biological complexity of a trait.

Candidate gene classification

In plants, over 1000 GWAS have been published in the last decade (H.-J. Liu & Yan, 2019). This highlights the rapid advancement in the field of crop research but also leads to questions about how to prioritize candidates for validation, given the relatively large effort and costs associated with such experiments (Mohammadi et al., 2020). Although advancements in transformation efficiency (Lowe et al., 2016) will certainly allow for more candidates to be validated, the production of high quality candidates continues to be of utmost importance. Prioritizing GWAS candidates by integrating gene co-expression networks as an additional source of evidence has been demonstrated to be a powerful approach to dissect grain elemental traits as well as insect resistance in maize (Badji et al., 2020; Schaefer et al., 2018). In our investigation, a multi-covariate GWAS approach meant to reduce type 1 and type 2 errors, together with a strict significance threshold, an integration with mutual co-expression analysis, and the use of composite phenotypes resulted in the selection of 7 high confidence candidate genes (Figure 2.4c). Two genes, in particular, ZmARF20 (Zm00001d015243) and

Zm00001d029527, are likely contenders for improving bioenergy and may warrant further experimental investigation in maize based on what is known about them in Arabidopsis (Baskin, 2001; Nakajima et al., 2004, 2006; Powers et al., 2019; Wilmoth et al., 2005). ZmARF20 (Zm00001d015243) was associated with the composite trait [GLU, SA, VBD], while Zm00001d029527 was associated with [PEN, VBD, SA].

In Arabidopsis, the ZmARF20 (Zm00001d015243) ortholog (At1G19220) encodes an auxin response factor (ARF) protein which transcriptionally activates auxin response elements (Wilmoth et al., 2005). Recent work has functionally characterized the dynamic localization of the At1G19220 ARF protein in Arabidopsis roots, and demonstrated that during cell growth it is active at the nucleus, whereas when the cell is not growing, it is relegated to the cytoplasm as a biomolecular condensate with reduced activity (Powers et al., 2019). If a similar mechanism occurs in maize stalks, a possible explanation for why ZmARF20 was associated with [GLU, SA, VBD] might be that the ZmARF20 protein from inbred lines at the extremes have contrasting preferential localization of to either the nucleus or cytoplasm where they are active, or inactive, respectively. Therefore, inbred lines with preferential nuclear localization might be more continuously responsive to auxin and would presumably grow larger more vascularized stalks. An additional explanation might be that inbreds with high [GLU, SA, VBD] more easily switch between the nucleus and cytoplasm, providing plasticity in auxin responsiveness to the stalk to grow when conditions are most favorable.

The top candidate associated with [PEN, VBD, SA], Zm00001d029527, does not encode a TF regulatory gene. However, the small 107 amino acid SPIRAL1-like protein which Zm00001d029527 encodes, likely acts in a regulatory capacity. In Arabidopsis, the role of this same protein is to to transversely orient cortical microtubules (Nakajima et al., 2004, 2006), which associate with cellulose synthase proteins complex to deposit cellulose to the cell wall and allow for normal anisotropic growth (Mizrachi et al., 2012). Interestingly, SPIRAL1 Arabidopsis mutants have disoriented cortical microtubules which result in helical growth patterns, but are not stunted (Nakajima et al., 2004, 2006). It is difficult to say for certain what the SPIRAL1-like protein is doing in maize stalks, or its connection to [PEN, VBD, SA]. Yet, we can speculate that SNP variation within Zm00001d029527 might be altering or truncating the peptide sequence of the SPIRAL1-like protein, which in turn, may result in modified microtubule organization, cell wall structure or distributions of cellulose in the stalk. If this were the case, further research on potentially a few inbreds with contrasting SNP states at the allele we associated with Zm00001d029527 could be evaluated using finer resolution phenotyping techniques to determine cell shape characteristics.

Here, we have presented a large-scale characterization and genetic dissection of traits relevant to biomass for bioenergy production. We focused our discussion on genes that were associated with composite phenotypes, but we also provide a valuable resource summarizing the genetic analysis of many individual anatomical and saccharification traits (Supplemental Table 2.2) The results from our study have the potential to be used for the improvement of plant biomass through breeding, or gene manipulation. The insights gained may also assist to guide future research concerning the underlying mechanisms which determine stalk biochemical and anatomical traits in maize.

2.6 Author Contributions

JR performed the data analysis and interpretation and composed the manuscript. MH collected experimental field data and processed stalk images to produce anatomical trait measurements. ND and SK conceived and supervised the experiments.

2.7 Acknowledgments

Computing resources and assistance were received from the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the OSG Consortium, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. The authors would like to thank the Michigan State Cell Wall Facility, particularly Nick Santoro and Cliff Foster for performing the glucose and pentose biochemical assays.

2.8 Funding

This work was funded by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). Sequencing was conducted by the DOE Joint Genome Institute supported by the DOE Office of Science DE-AC02-05CH11231. Additional support from the National Science Foundation under Grant No. 1733633. JR was supported by a University of Wisconsin-Madison SciMed GRS fellowship. Support for this fellowship is provided by the Graduate School, part of the Office of Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison, with funding from the Wisconsin Alumni Research Foundation.

2.9 Tables

Table 2.1: Variance decomposition and entry mean narrow-sense heritability (h^2) estimates for stalk anatomical and saccharification traits evaluated on the WiDiv. Year variance (σ_E^2); Genotypic variance (σ_G^2); Replication within year variance (σ_R^2); Genotype-by-year variance (σ_{GxE}^2) and Residual error variance (σ_ε^2).

Trait†	σ_E^2	σ_G^2	σ_R^2	σ_{GxE}^2	σ_ε^2	h^2
GLU	5.80 x 10^{-1**}	3.25 x 10^{0***}	-	6.65 x 10^{-1***}	2.42 x 10^{0}	0.78
PEN	-	6.93 x 10^{-1***}	1.03 x 10^{-3}	3.59 x 10^{-1***}	7.17 x 10^{-1}	0.66
SDL	-	5.59x 10^{-2***}	5.30 x 10^{-4***}	-	3.15 x 10^{-2}	0.88
SDS	-	4.30 x 10^{-2***}	2.09 x 10^{-4*}	1.27 x 10^{-16}	2.47 x 10^{-2}	0.87
RT	-	6.28 x 10^{-3***}	-	-	1.86 x 10^{-2}	0.58
VBA	-	2.17 x 10^{-8***}	6.80 x 10^{-11}	-	3.22 x 10^{-8}	0.73
VBD	1.12 x 10^{-2}	8.44 x 10^{1***}	-	1.27 x 10^{-12}	6.70 x 10^{1}	0.83
SA	-	4.61 x 10^{-1***}	3.07 x 10^{-3**}	-	2.72 x 10^{-1}	0.87
PA	4.95 x 10^{-15}	1.85 x 10^{-1***}	1.74 x 10^{-3*}	3.06 x 10^{-14}	2.41 x 10^{-1}	0.75
TB	-	4.08 x 10^{2***}	2.71 x 10^{0*}	-	4.57 x 10^{2}	0.78
TBA	-	4.98 x 10^{-4***}	-	7.92 x 10^{-18}	6.95 x 10^{-4}	0.74
RA	-	8.80 x 10^{-2***}	2.65 x 10^{-4}	-	1.36 x 10^{-1}	0.72
GDDP	1.33 x 10^{3*}	1.09 x 10^{4***}	7.01 x 10^{1***}	1.19 x 10^{2*}	8.95 x 10^{2}	0.97
GDDS	1.00 x 10^{3}	1.23 x 10^{4***}	1.02 x 10^{2***}	1.11 x 10^{2}	1.08 x 10^{3}	0.97

† GLU, glucose (%/mg DM); PEN, pentose (%/mg DM); SDL, stalk diameter long-axis (cm); SDS, stalk diameter short-axis (cm); RT, rind thickness (cm); VBA, average vascular bundle area (cm^2); VBD, vascular bundle density (cm^{-2}); SA, stalk area (cm^2); PA, pith area (cm^2); TB, total bundles (counts); TBA, total bundle area (cm^2); RA, rind area (cm^2); GDDP, cumulative growing degree days from planting to pollen shed (°F); GDDS, cumulative growing degree days to ear silking (°F).Significance codes are from likelihood ratio test: Pr(>Chisq) <0.001 `***`; <0.01 `**`; <0.05 `*`; not significant ` `; `-` indicates zero variance estimate.

2.10 Figures

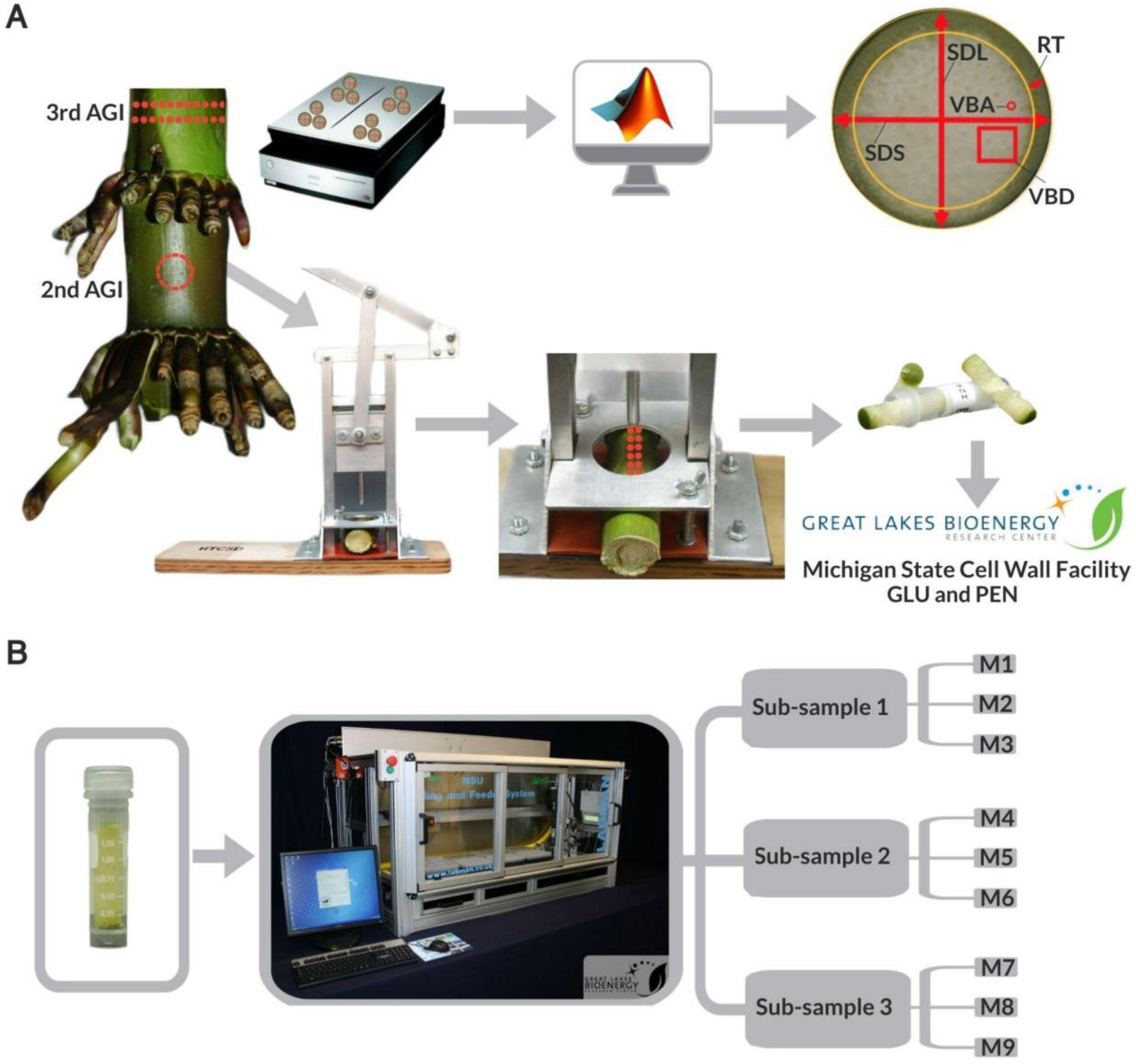


Figure 2.1: Dissection and processing steps following sample collection of the second and third above-ground stalk internodes (AGI). (A) A cross-sectional slice and a cylindrical core were dissected from each stalk sample. Anatomical traits were quantified by imaging the third AGI using a flat-bed scanner and processing the images in MATLAB. The features extracted included stalk diameter short-axis (SDS), stalk diameter long-axis (SDL), average vascular bundle area (VBA), vascular bundle density (VBD), and rind thickness (RT). Cylindrical cores were isolated using a high-throughput core sampling device. Glucose (GLU) and pentose (PEN) biochemical assays were conducted by the Michigan State Cell Wall Facility. (B) Each stalk core was processed through the iWALL system. The platform ground and divided each core sample into three sub-samples. Assays were conducted in triplicate, resulting in a total of nine technical replicate measurements per core sample.

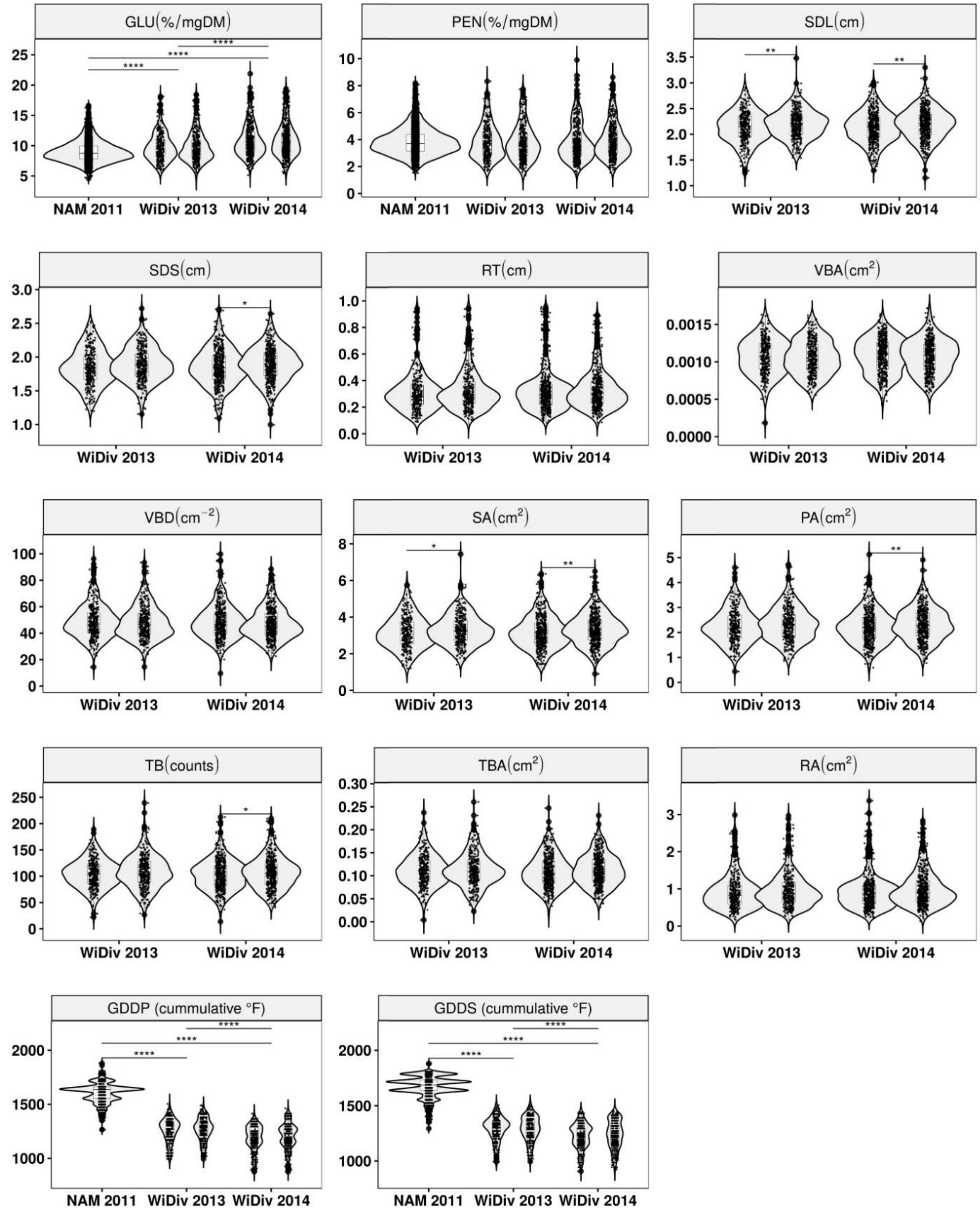


Figure 2.2: Phenotypic distributions separated by replication, year evaluated, and population. Depicted are plot-level values for each trait. Significance codes are from two-sided T-tests comparing pairwise differences between groups, where Pr(>T) <0.0001 `****`; <0.001 `***` ; <0.01 `**` ; <0.05 `*` ; not significant ` `.

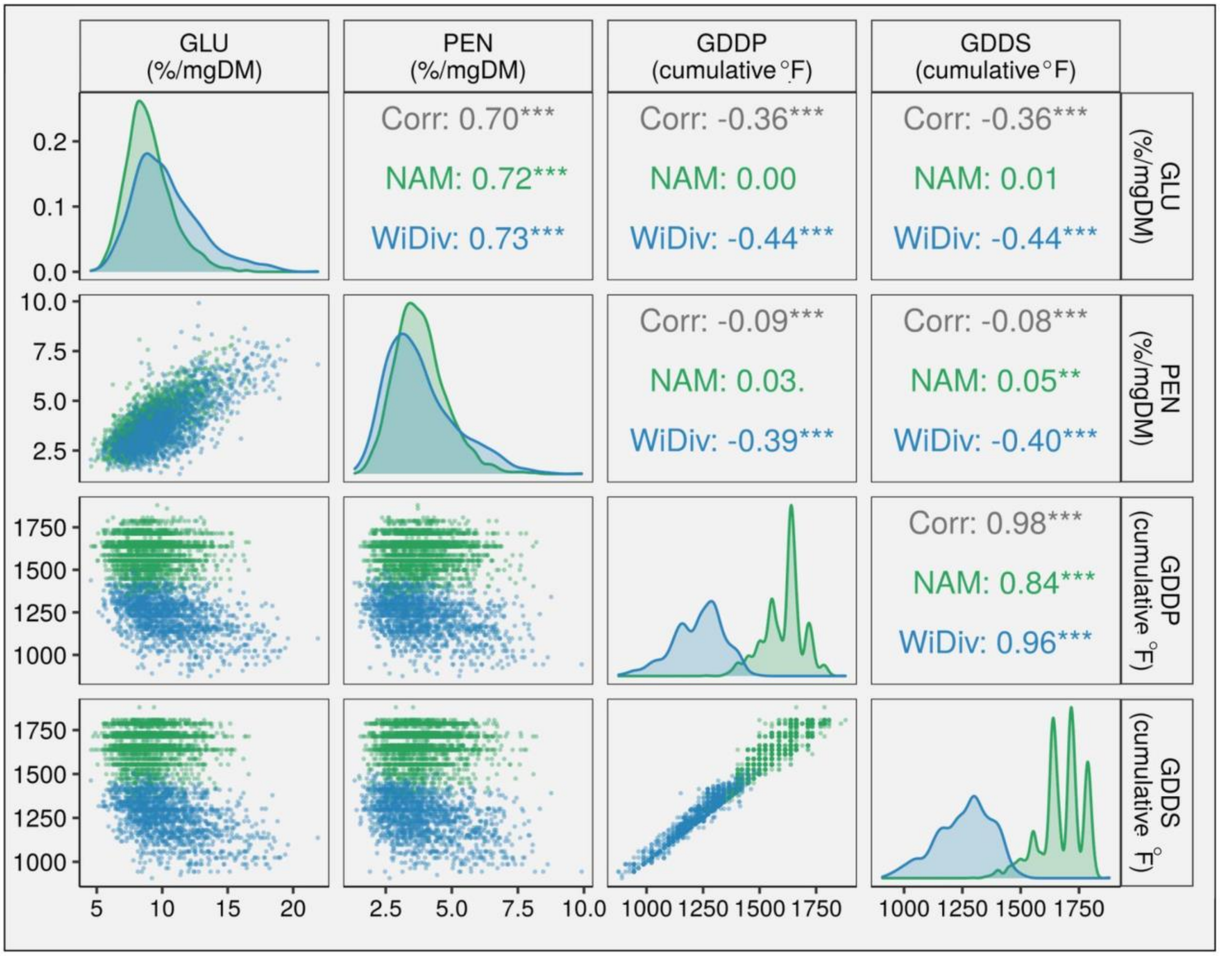


Figure 2.3: Pearson correlation matrix of saccharification and flowering time traits. Each data point represents a single plot level measurement. In the upper triangle, correlations are shown for the combined analysis of WiDiv and NAM data in grey. The correlations separated by population are depicted in green for the NAM and blue for the WiDiv. Significance codes are from two-sided T-test: Pr(>T) <0.001 `***`; <0.01 `**`; <0.05 `*`; not significant ` `.

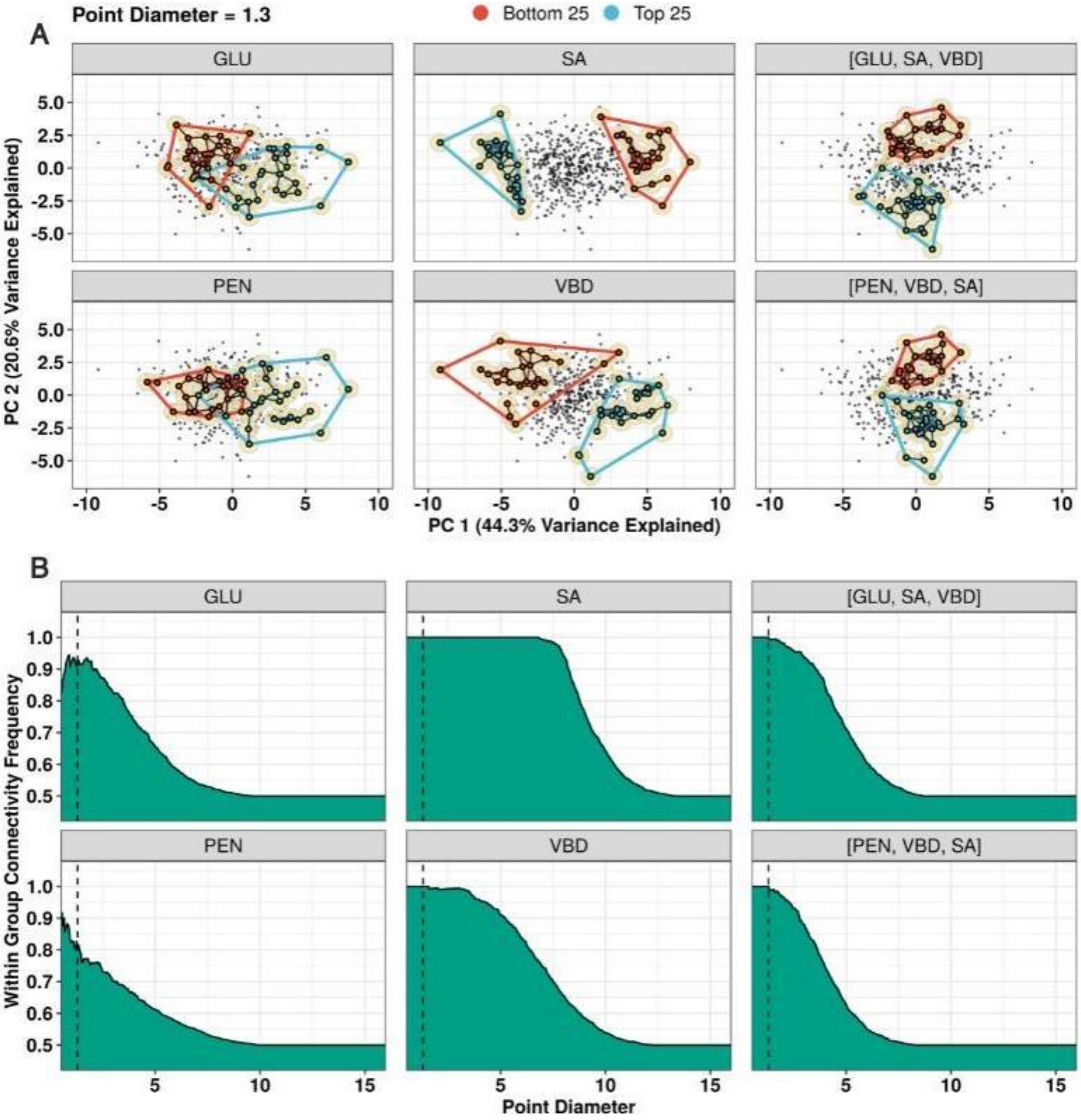


Figure 2.4: Clustering and connectivity of WiDiv lines expressing extreme (top 25 and bottom 25) saccharification (GLU and PEN), a subset of anatomical traits (SA and VBD), and composite saccharification and anatomical phenotypes (enclosed in brackets). (A) Biplots of the first two principal components calculated from all saccharification and anatomical phenotypes together. For each phenotype, the convex hull (outermost polygon) of each phenotypic extreme is shown, along with the connectivity within (red or blue line segments) and between (black line segments) extreme groups. Plotted is the static frame when using a point diameter of 1.3. (B) The within group connectivity frequency as a function of increasing point diameter size. The dashed horizontal line corresponds to the point diameter of 1.3 shown in (A).

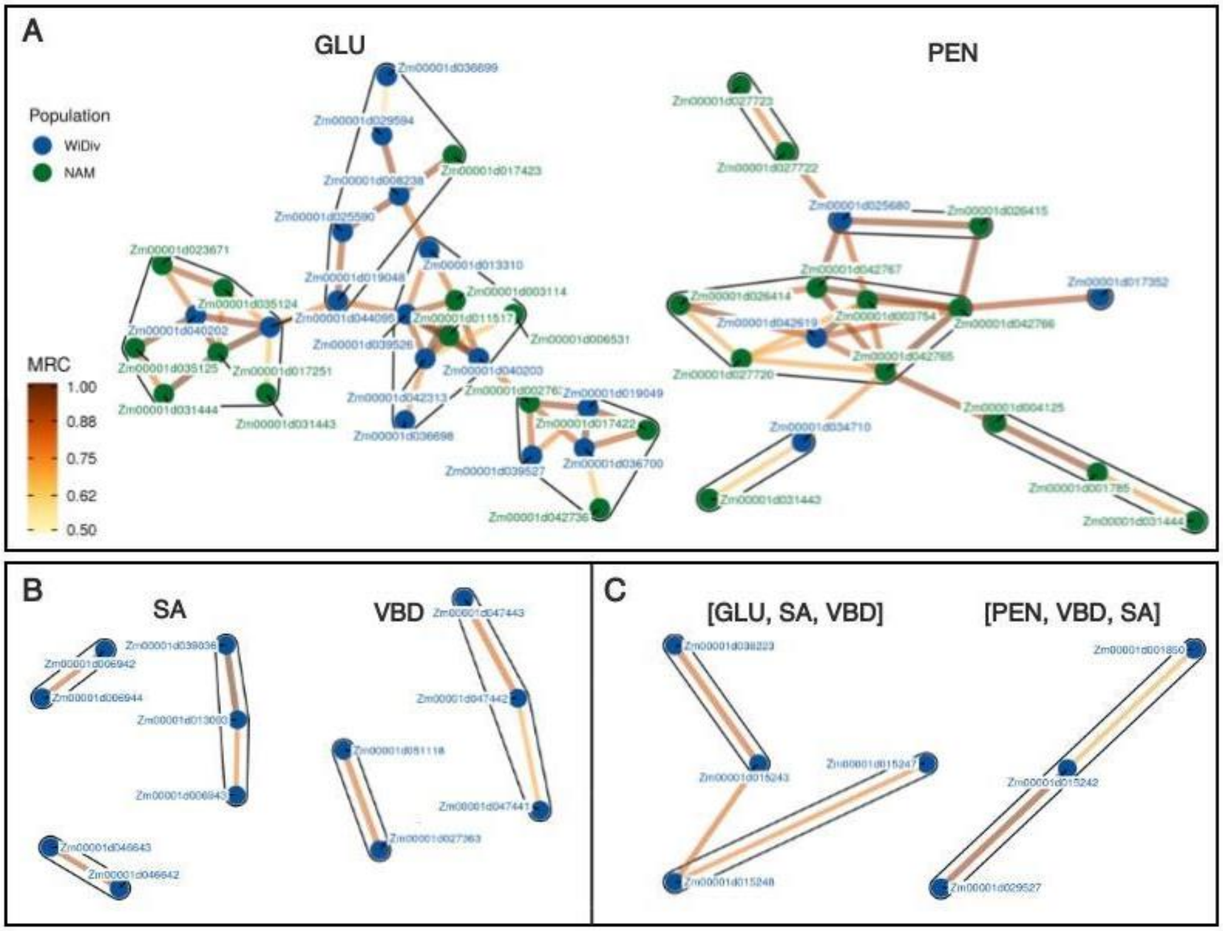


Figure 2.5: Candidate genes (nodes) identified by GWAS and co-expression for saccharification and a subset of stalk anatomical and composite phenotypes. The edges forming connections between genes in each panel are shaded based on the strength of their mutual rank co-expression coefficient (MRC) from gene expression at the seedling stage. Groupings surrounded by a solid line within each panel gene cliques formed based on edge-betweenness centrality scores. In (A) are candidate genes identified for the two stalk saccharification traits collected on the WiDiv and NAM populations. In (B) and (C) are candidates for traits measured on the WiDiv population alone. These represent two anatomical traits alone (B), and as a composite (C) with the saccharification traits.

2.11 Supplemental Tables

Supplemental Table 2.1: Inbred lines belonging to the WiDiv and NAM populations evaluated as part of this study and the corresponding phenotypic data on a per-plot basis (Deposited as a separate file with this thesis).

Supplemental Table 2.2: GWAS SNP-trait associations for the WiDiv and NAM populations (Deposited as a separate file with this thesis).

2.12 Supplemental Files

Supplemental File 2.1: Animation of Figure 2.3, depicting the relationship between increasing point size diameter (yellow disks) and the connectivity within (red and blue line segments) and between (black line segments) phenotypic extreme groups (Deposited as a separate file with this thesis).

2.13 Supplemental Figures

Supplemental Figure 2.1: Principal component and kinship analysis using LD pruned SNP markers from the (A) WiDiv and (B) NAM. The values along the diagonal of the principal component biplot matrices correspond to the variance explained by each principal component. For each population, the Pearson correlation between the phenotypes measured and the top five principal components derived from the SNP data are shown. Each subpopulation (WiDiv) or bi-parental RIL family (NAM) is represented by a different color in the respective plot. Significance codes in the correlation heatmap are from a two-sided T-test: Pr(>T) <0.001 `***`; <0.01 `**`; <0.05 `*`; not significant ` `. (Deposited as a separate file with this thesis).

Supplemental Figure 2.2: (A) Multi-covariate model GWAS approach used for mapping SNP-trait associations. (B) The GWAS approach applied to flowering time and subsequent identification of two previously cloned genes controlling flowering time in maize. (Deposited as a separate file with this thesis).

Supplemental Figure 2.3: Pearson correlation matrix of saccharification, anatomical, and flowering time traits collected on the WiDiv population. Each data point represents a single plot level measurement. Significance codes are from two-sided T-test: Pr(>T) <0.001 `***`; <0.01 `**`; <0.05 `*`; not significant ` `. (Deposited as a separate file with this thesis).

Supplemental Figure 2.4: Spearman rank correlations of stalk saccharification and anatomical traits between years. Correlations were calculated from plot means by year. All correlations are significant (Pr>T <0.001). (Deposited as a separate file with this thesis).

Supplemental Figure 2.5: The first two principal components calculated from BLUPs of all saccharification, and anatomical phenotypes collected on the WiDiv population. (Deposited as a separate file with this thesis).

Supplemental Figure 2.6: (A) GWAS model comparison of true and false positive rates using simulated data. (B) Area under the smoothed receiver operator characteristic (ROC) curve shown in (A). Significance lettering is based on lsmeans with an alpha of 0.05. (Deposited as a separate file with this thesis).

Supplemental Figure 2.7: (A-B) Quantile-Quantile and (C-D) manhattan plots for GWAS conducted on the WiDiv (A & C) and NAM (B & D). The GWAS models fit for each trait included a naive single marker T-test, and a series of FarmCPU models accounting for kinship correction (K), and sequentially increasing principal components (PCs) up to 5PCs to correct for population structure (Q). SNP markers highlighted in (C & D) surpassed the respectively labeled threshold, where red is the Bonferroni threshold (0.05/M), yellow is the permutation threshold (1000x per trait), and green is the suggestive threshold (1/M). (Deposited as a separate file with this thesis).

2.14 References

Amézquita, E. J., Quigley, M. Y., Ophelders, T., Munch, E., & Chitwood, D. H. (2020). The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics*, *249*(7), 816–833. https://doi.org/10.1002/dvdy.175

Badji, A., Kwemoi, D. B., Machida, L., Okii, D., Mwila, N., Agbahoungba, S., Kumi, F., Ibanda, A., Bararyenya, A., Solemanegy, M., Odong, T., Wasswa, P., Otim, M., Asea, G., Ochwo-Ssemakula, M., Talwana, H., Kyamanywa, S., & Rubaihayo, P. (2020). Genetic Basis of Maize Resistance to Multiple Insect Pests: Integrated Genome-Wide Comparative Mapping and Candidate Gene Prioritization. *Genes*, *11*(6), 689. https://doi.org/10.3390/genes11060689

Barrière, Y., Courtial, A., Chateigner-Boutin, A. L., Denoue, D., & Grima-Pettenati, J. (2015). Breeding maize for silage and biofuel production, an illustration of a step forward with the genome sequence. *Plant Science*, *242*, 310–329. https://doi.org/10.1016/j.plantsci.2015.08.007

Barros-Rios, J., Santiago, R., Malvar, R. A., & Jung, H. J. G. (2012). Chemical composition and cell wall polysaccharide degradability of pith and rind tissues from mature maize internodes. *Animal Feed Science and Technology*, *172*(3–4), 226–236. https://doi.org/10.1016/j.anifeedsci.2012.01.005

Baskin, T. I. (2001). On the alignment of cellulose microfibrils by cortical microtubules: A review and a model. *Protoplasma*, *215*(1–4), 150–171. https://doi.org/10.1007/BF01280311

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

Bichot, A., Delgenès, J.-P., Méchin, V., Carrère, H., Bernet, N., & García-Bernet, D. (2018). Understanding biomass recalcitrance in grasses for their efficient utilization as biorefinery feedstock. *Reviews in Environmental Science and Bio/Technology*, *17*(4), 707–748. https://doi.org/10.1007/s11157-018-9485-y

Brigham, C. (2018). Chapter 3.22 - Biopolymers: Biodegradable Alternatives to Traditional Plastics. In B. Török & T. Dransfield (Eds.), *Green Chemistry* (pp. 753–770). Elsevier. https://doi.org/10.1016/B978-0-12-809270-5.00027-3

Broad Institute. (2019). *Picard toolkit*. https://broadinstitute.github.io/picard/

Brunson, J. C., Wadhwa, R., & Scott, J. (2021). *ggtda: Ggplot2 Extension to Visualize Topological Persistence*. https://rrrlw.github.io/ggtda/

Chazal, F., & Michel, B. (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, *4*, 108. https://doi.org/10.3389/frai.2021.667963

Chen, F., & Dixon, R. A. (2007). Lignin modification improves fermentable sugar yields for biofuel production. *Nature Biotechnology*, *25*(7), 759–761. https://doi.org/10.1038/nbt1316

Deschatelets, L., & Yu, E. K. C. (1986). A simple pentose assay for biomass conversion studies. *Applied Microbiology and Biotechnology*, *24*(5), 379–385. https://doi.org/10.1007/BF00294594

Duvick, D. N. (2005). The Contribution of Breeding to Yield Advances in maize (Zea mays L.). In *Advances in Agronomy* (Vol. 86, pp. 83–145). Elsevier. https://doi.org/10.1016/S0065-2113(05)86002-X

Duvick, D. N., Smith, J. S. C., & Cooper, M. (2003). Long-Term Selection in a Commercial Hybrid Maize Breeding Program. In *Plant Breeding Reviews* (pp. 109–151). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470650288.ch4

Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2000). Topological persistence and simplification. *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 454–463. https://doi.org/10.1109/SFCS.2000.892133

El Bassam, N. (2010). Energy Crops Guide. In *Handbook of bioenergy crops: A complete reference to species, development and applications* (1st ed., pp. 93–396). Earthscan.

Ensembl Plants v48. (2019). *Ensembl Plants*. https://plants.ensembl.org/index.html

Fernández-Sandoval, M. T., Galíndez-Mayer, J., Bolívar, F., Gosset, G., Ramírez, O. T., & Martinez, A. (2019). Xylose–glucose co-fermentation to ethanol by Escherichia coli strain MS04 using single- and two-stage continuous cultures under micro-aerated conditions. *Microbial Cell Factories*, *18*(1), 145. https://doi.org/10.1186/s12934-019-1191-0

Fornalé, S., Capellades, M., Encina, A., Wang, K., Irar, S., Lapierre, C., Ruel, K., Joseleau, J. P., Berenguer, J., Puigdomènech, P., Rigau, J., & Caparrós-Ruiz, D. (2012). Altered lignin biosynthesis improves cellulosic bioethanol production in transgenic maize plants down-regulated for cinnamyl alcohol dehydrogenase. *Molecular Plant*, *5*(4), 817–830. https://doi.org/10.1093/mp/ssr097

Fu, C., Xiao, X., Xi, Y., Ge, Y., Chen, F., Bouton, J., Dixon, R. A., & Wang, Z. Y. (2011). Downregulation of Cinnamyl Alcohol Dehydrogenase (CAD) Leads to Improved Saccharification Efficiency in Switchgrass. *Bioenergy Research*, *4*(3), 153–164. https://doi.org/10.1007/s12155-010-9109-z

Gachomo, E. W., Jimenez-Lopez, J. C., Baptiste, L. J., & Kotchoni, S. O. (2014). GIGANTUS1 (GTS1), a member of Transducin/WD40 protein superfamily, controls seed germination, growth and biomass accumulation through ribosome-biogenesis protein interactions in Arabidopsis thaliana. *BMC Plant Biology*, *14*, 37. https://doi.org/10.1186/1471-2229-14-37

Galbe, M., & Wallberg, O. (2019). Pretreatment for biorefineries: A review of common methods for efficient utilisation of lignocellulosic materials. *Biotechnology for Biofuels*, *12*(1), 294. https://doi.org/10.1186/s13068-019-1634-1

Gilmore, E. C., & Rogers, J. S. (1958). Heat Units as a Method of Measuring Maturity in Corn1. *Agronomy Journal*, *50*(10), 611–615. https://doi.org/10.2134/agronj1958.00021962005000100014x

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Golecha, R., & Gan, J. (2016). Effects of corn stover year-to-year supply variability and market structure on biomass utilization and cost. *Renewable and Sustainable Energy Reviews*, *57*, 34–44. https://doi.org/10.1016/j.rser.2015.12.075

Gomez, F. E., Carvalho, G., Shi, F., Muliana, A. H., & Rooney, W. L. (2018). High throughput phenotyping of morpho-anatomical stem properties using X-ray computed tomography in sorghum. *Plant Methods*, *14*(1), 1–13. https://doi.org/10.1186/s13007-018-0326-3

Gray, J., Bevan, M., Brutnell, T., Buell, C. R., Cone, K., Hake, S., Jackson, D., Kellogg, E., Lawrence, C., McCouch, S., Mockler, T., Moose, S., Paterson, A., Peterson, T., Rokshar, D., Souza, G. M., Springer, N., Stein, N., Timmermans, M., … Grotewold, E. (2009). A Recommendation for Naming Transcription Factor Proteins in the Grasses. *Plant Physiology*, *149*(1), 4–6. https://doi.org/10.1104/pp.108.128504

Heckwolf, S., Heckwolf, M., Kaeppler, S. M., de Leon, N., & Spalding, E. P. (2015). Image analysis of anatomical traits in stalk transections of maize and other grasses. *Plant Methods*, *11*(26), 26. https://doi.org/10.1186/s13007-015-0070-x

Hoang Nguyen Tran, P., Ko, J. K., Gong, G., Um, Y., & Lee, S.-M. (2020). Improved simultaneous co-fermentation of glucose and xylose by Saccharomyces cerevisiae for efficient lignocellulosic biorefinery. *Biotechnology for Biofuels*, *13*(1), 12. https://doi.org/10.1186/s13068-019-1641-2

Jung, H. G., Morrison, T. A., & Buxton, D. R. (1998). Degradability of cell-wall polysaccharides in maize internodes during stalk development. *Crop Science*, *38*(4), 1047–1051. https://doi.org/10.2135/cropsci1998.0011183X003800040027x

Kang, X., Kirui, A., Dickwella Widanage, M. C., Mentink-Vigier, F., Cosgrove, D. J., & Wang, T. (2019). Lignin-polysaccharide interactions in plant secondary cell walls revealed by solid-state NMR. *Nature Communications*, *10*. https://doi.org/10.1038/s41467-018-08252-0

Khare, S. K., Pandey, A., & Larroche, C. (2015). Current perspectives in enzymatic saccharification of lignocellulosic biomass. *Biochemical Engineering Journal*, *102*, 38–44. https://doi.org/10.1016/j.bej.2015.02.033

Kim, J. S., Lee, Y. Y., & Kim, T. H. (2016). A review on alkaline pretreatment technology for bioconversion of lignocellulosic biomass. *Bioresource Technology*, *199*, 42–48. https://doi.org/10.1016/j.biortech.2015.08.085

Kumar, A., Kumar, V., Kumar, A., Antonio, F., Antunes, F., & Silvério, S. (2018). Bioresource Technology The path forward for lignocellulose biorefineries: Bottlenecks, solutions, and perspective on commercialization. *Bioresource Technology*, *264*(June), 370–381. https://doi.org/10.1016/j.biortech.2018.06.004

Kusmec, A., Srinivasan, S., Nettleton, D., & Schnable, P. S. (2017). Distinct genetic architectures for phenotype means and plasticities in Zea mays. *Nature Plants*, *3*(9), 715–723. https://doi.org/10.1038/s41477-017-0007-7

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Lenth, R. V. (2016). Least-Squares Means: The *R* Package **lsmeans**. *Journal of Statistical Software*, *69*(1). https://doi.org/10.18637/jss.v069.i01

Li, M., Heckwolf, M., Crowe, J. D., Williams, D. L., Magee, T. D., Kaeppler, S. M., De Leon, N., & Hodge, D. B. (2015). Cell-wall properties contributing to improved deconstruction by alkaline pre-treatment and enzymatic hydrolysis in diverse maize (Zea mays L.) lines. *Journal of Experimental Botany*, *66*(14), 4305–4315. https://doi.org/10.1093/jxb/erv016

Liang, Y., Liu, Q., Wang, X., Huang, C., Xu, G., Hey, S., Lin, H. Y., Li, C., Xu, D., Wu, L., Wang, C., Wu, W., Xia, J., Han, X., Lu, S., Lai, J., Song, W., Schnable, P. S., & Tian, F. (2019). ZmMADS69 functions as a flowering activator through the ZmRap2.7-ZCN8 regulatory module and contributes to maize flowering time adaptation. *New Phytologist*, *221*(4), 2335–2347. https://doi.org/10.1111/nph.15512

Liu, H.-J., & Yan, J. (2019). Crop genome-wide association study: A harvest of biological relevance. *The Plant Journal*, *97*(1), 8–18. https://doi.org/10.1111/tpj.14139

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. *BMC Genetics*, *13*(1), 100–100. https://doi.org/10.1186/1471-2156-13-100

López-Malvar, A., Butron, A., Malvar, R. A., McQueen-Mason, S. J., Faas, L., Gómez, L. D., Revilla, P., Figueroa-Garrido, D. J., & Santiago, R. (2021). Association mapping for maize stover yield and saccharification efficiency using a multiparent advanced generation intercross (MAGIC) population. *Scientific Reports*, *11*, 3425. https://doi.org/10.1038/s41598-021-83107-1

Lorenzana, R. E., Lewis, M. F., Jung, H. J. G., & Bernardo, R. (2010). Quantitative trait loci and trait correlations for maize stover cell wall composition and glucose release for cellulosic ethanol. *Crop Science*, *50*(2), 541–555. https://doi.org/10.2135/cropsci2009.04.0182

Lowe, K., Wu, E., Wang, N., Hoerster, G., Hastings, C., Cho, M.-J., Scelonge, C., Lenderts, B., Chamberlin, M., Cushatt, J., Wang, L., Ryan, L., Khan, T., Chow-Yiu, J., Hua, W., Yu, M., Banh, J., Bao, Z., Brink, K., … Gordon-Kamm, W. (2016). Morphogenic Regulators *Baby boom* and *Wuschel* Improve Monocot Transformation. *The Plant Cell*, *28*(9), 1998–2015. https://doi.org/10.1105/tpc.16.00124

Matos, D. A., Whitney, I. P., Harrington, M. J., & Hazen, S. P. (2013). Cell walls and the developmental anatomy of the Brachypodium distachyon stem internode. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0080640

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., Robin Buell, C., de Leon, N., & Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, *19*(1), 1–17. https://doi.org/10.1186/s12870-019-1653-x

Meyer, H. V., & Birney, E. (2018). PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics*, *34*(17), 2951–2956. https://doi.org/10.1093/bioinformatics/bty197

Mizrachi, E., Mansfield, S. D., & Myburg, A. A. (2012). Cellulose factories: Advancing bioenergy production from forest trees. *New Phytologist*, *194*(1), 54–62. https://doi.org/10.1111/j.1469-8137.2011.03971.x

Mohammadi, M., Xavier, A., Beckett, T., Beyer, S., Chen, L., Chikssa, H., Cross, V., Freitas Moreira, F., French, E., Gaire, R., Griebel, S., Lopez, M. A., Prather, S., Russell, B., & Wang, W. (2020). Identification, deployment, and transferability of quantitative trait loci from genome-wide association studies in plants. *Current Plant Biology*, *24*, 100145. https://doi.org/10.1016/j.cpb.2020.100145

Morrison, T. A., Jung, H. G., Buxton, D. R., & Hatfield, R. D. (1998). Cell-Wall Composition of Maize Internodes of Varying Maturity. *Crop Science*, *38*(2), 455–460. https://doi.org/10.2135/cropsci1998.0011183X003800020031x

Munaiz, E. D., Albrecht, K. A., & Ordas, B. (2021). Genetic Diversity for Dual Use Maize: Grain and Second-Generation Biofuel. *Agronomy*, *11*(2), 230. https://doi.org/10.3390/agronomy11020230

Muttoni, G., Johnson, J. M., Santoro, N., Rhiner, C. J., von Mogel, K. J. H., Kaeppler, S. M., & de Leon, N. (2012). A high-throughput core sampling device for the evaluation of maize stalk composition. *Biotechnology for Biofuels*, *5*(1), 27–27. https://doi.org/10.1186/1754-6834-5-27

Nakajima, K., Furutani, I., Tachimoto, H., Matsubara, H., & Hashimoto, T. (2004). SPIRAL1 encodes a plant-specific microtubule-localized protein required for directional control of rapidly expanding Arabidopsis cells. *The Plant Cell*, *16*(5), 1178–1190. https://doi.org/10.1105/tpc.017830

Nakajima, K., Kawamura, T., & Hashimoto, T. (2006). Role of the SPIRAL1 Gene Family in Anisotropic Growth of Arabidopsis thaliana. *Plant and Cell Physiology*, *47*(4), 513–522. https://doi.org/10.1093/pcp/pcj020

Obayashi, T., & Kinoshita, K. (2009). Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. *DNA Research*, *16*(5), 249–260. https://doi.org/10.1093/dnares/dsp016

Pedersen, J. F., Vogel, K. P., & Funnell, D. L. (2005). Impact of Reduced Lignin on Plant Fitness. *Crop Science*, *45*(3), 812–819. https://doi.org/10.2135/cropsci2004.0155

Penning, B. W., Hunter 3rd, C. T., Tayengwa, R., Eveland, A. L., Dugard, C. K., Olek, A. T., Vermerris, W., Koch, K. E., McCarty, D. R., Davis, M. F., Thomas, S. R., McCann, M. C., & Carpita, N. C. (2009). Genetic resources for maize cell wall biology. *Plant Physiol*, *151*(4), 1703–1728. https://doi.org/10.1104/pp.109.136804

Penning, B. W., Sykes, R. W., Babcock, N. C., Dugard, C. K., Held, M. A., Klimek, J. F., Shreve, J. T., Fowler, M., Ziebell, A., Davis, M. F., Decker, S. R., Turner, G. B., Mosier, N. S., Springer, N. M., Thimmapuram, J., Weil, C. F., Mccann, M. C., & Carpita, N. C. (2014). Genetic Determinants for Enzymatic Digestion of Lignocellulosic Biomass are Independent of Those for Lignin abundance in a maize Recombinant Inbred population. *Plant Physiology*, *165*(4), 1475–1487. https://doi.org/10.1104/pp.114.242446

Powers, S. K., Holehouse, A. S., Korasick, D. A., Schreiber, K. H., Clark, N., Jing, H., Emenecker, R., Han, S., Tycksen, E., Hwang, I., Sozzani, R., Jez, J. M., Pappu, R. V., & Strader, L. C. (2019). Nucleo-cytoplasmic partitioning of ARF proteins controls auxin responses in Arabidopsis thaliana. *Molecular Cell*, *76*(1), 177-190.e5. https://doi.org/10.1016/j.molcel.2019.06.044

Pradhan Mitra, P., & Loqué, D. (2014). Histochemical staining of Arabidopsis thaliana secondary cell wall elements. *Journal of Visualized Experiments*, *87*, 1–11. https://doi.org/10.3791/51381

Price, M. N., & Arkin, A. P. (2017). PaperBLAST: Text Mining Papers for Information about Homologs. *MSystems*, *2*(4), e00039-17. https://doi.org/10.1128/mSystems.00039-17

Prive, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics*, *34*(16), 2781–2787. https://doi.org/10.1093/bioinformatics/bty185

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Ruan, Z., Wang, X., Liu, Y., & Liao, W. (2019). Chapter 3—Corn. In Z. Pan, R. Zhang, & S. Zicari (Eds.), *Integrated Processing Technologies for Food and Agricultural By-Products* (pp. 59–72). Academic Press. https://doi.org/10.1016/B978-0-12-814138-0.00003-4

Santoro, N., Cantu, S. L., Tornqvist, C.-E., Falbel, T. G., Bolivar, J. L., Patterson, S. E., Pauly, M., & Walton, J. D. (2010). A High-Throughput Platform for Screening Milligram Quantities of Plant Biomass for Lignocellulose Digestibility. *BioEnergy Research*, *3*(1), 93–102. https://doi.org/10.1007/s12155-009-9074-6

Schaefer, R. J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., & Myers, C. L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. *The Plant Cell*, *30*(12), 2922–2942. https://doi.org/10.1105/tpc.18.00299

Scheller, H. V., & Ulvskov, P. (2010). Hemicelluloses. *Annual Review of Plant Biology*, *61*(1), 263–289. https://doi.org/10.1146/annurev-arplant-042809-112315

Shane, M. W., McCully, M. E., & Canny, M. J. (2000). The vascular system of maize stems revisited: Implications for water transport and xylem safety. *Annals of Botany*, *86*(2), 245–258. https://doi.org/10.1006/anbo.2000.1171

Slack, C. R., & Hatch, M. D. (1967). Comparative studies on the activity of carboxylases and other enzymes in relation to the new pathway of photosynthetic carbon dioxide fixation in tropical grasses. *The Biochemical Journal*, *103*(3), 660–665. https://doi.org/10.1042/bj1030660

Stanger, T. F., & Lauer, J. G. (2007). Corn Stalk Response to Plant Population and the Bt–European Corn Borer Trait. *Agronomy Journal*, *99*(3), 657–664. https://doi.org/10.2134/agronj2006.0079

The ImageMagick Development Team. (2021). *ImageMagick* (7.0.10) [Computer software]. https://imagemagick.org

Tobimatsu, Y., Wouwer, D. V. de, Allen, E., Kumpf, R., Vanholme, B., Boerjan, W., & Ralph, J. (2014). A click chemistry strategy for visualization of plant cell wall lignification. *Chem. Commun.*, *50*(82), 12262–12265. https://doi.org/10.1039/C4CC04692G

Torres, A. F., Slegers, P. M., Noordam-Boot, C. M. M., Dolstra, O., Vlaswinkel, L., Van Boxtel, A. J. B., Visser, R. G. F., & Trindade, L. M. (2016). Maize feedstocks with improved digestibility reduce the costs and environmental impacts of biomass pretreatment and saccharification. *Biotechnology for Biofuels*, *9*(1), 1–15. https://doi.org/10.1186/s13068-016-0479-0

U.S. Congress. (2007). Energy independence and security act. *Public Law*, *110–140*, 1492–1801.

U.S. Department of Energy. (2011). *U.S. Billion-Ton Update: Biomass Supply for a Bioenergy and Bioproducts Industry. R.D. Perlack and B.J. Stokes (Leads), ORNL/TM-2011/224. Oak Ridge National Laboratory, Oak Ridge, TN. 227p.* 235.

U.S. Energy Information Administration. (2021). *Annual Energy Outlook 2021* (pp. 1–33). https://www.eia.gov/outlooks/aeo/

Wilmoth, J. C., Wang, S., Tiwari, S. B., Joshi, A. D., Hagen, G., Guilfoyle, T. J., Alonso, J. M., Ecker, J. R., & Reed, J. W. (2005). NPH4/ARF7 and ARF19 promote leaf expansion and auxin-induced lateral root formation. *The Plant Journal*, *43*(1), 118–130. https://doi.org/10.1111/j.1365-313X.2005.02432.x

Wilson, J. R., & Hatfield, R. D. (1997). Structural and chemical changes of cell wall types during stem development: Consequences for fibre degradation by rumen microflora. *Australian Journal of Agriculture Research*, *48*(2), 165–180. https://doi.org/10.1071/A96051

Wilson, J. R., & Mertens, D. R. (1995). Cell Wall Accessibility and Cell Structure Limitations to Microbial Digestion of Forage. *Crop Science*, *35*(1), 251–259. https://doi.org/10.2135/cropsci1995.0011183x003500010046x

Yilmaz, A., Nishiyama, M. Y., Jr., Fuentes, B. G., Souza, G. M., Janies, D., Gray, J., & Grotewold, E. (2009). GRASSIUS: A Platform for Comparative Regulatory Genomics across the Grasses. *Plant Physiology*, *149*(1), 171–180. https://doi.org/10.1104/pp.108.128579

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2020.10.007

Yoon, J., Choi, H., & An, G. (2015). Roles of lignin biosynthesis and regulatory genes in plant development. *Journal of Integrative Plant Biology*, *57*(11), 902–912. https://doi.org/10.1111/jipb.12422

Yu, J., Holland, J. B., McMullen, M. D., & Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, *178*(1), 539–551. https://doi.org/10.1534/genetics.107.074245

Zhang, Y., Du, J., Wang, J., Ma, L., Lu, X., Pan, X., Guo, X., & Zhao, C. (2018). High-throughput micro-phenotyping measurements applied to assess stalk lodging in maize (Zea mays L.). *Biological Research*, *51*(1), 1–14. https://doi.org/10.1186/s40659-018-0190-7

Zhou, P., & Springer, N. M. (2020). *Data for: Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions* [Data set]. https://doi.org/10.13020/p3g0-3170

Chapter 3: Normalizing and correcting variable and complex LC–MS metabolomic data with the R package pseudoDrift

Authors:

Jonas Rodriguez[1], Lina Gomez-Cano[2], Natalia de Leon[1], Erich Grotewold[2]

Affiliations:

[1] Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA

[2] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, 48824, USA

Publication information: This article was published in the journal *Metabolites* on May 12, 2022.
DOI: https://doi.org/10.3390/metabo12050435

3.1 Abstract

In biological research domains, liquid chromatography–mass spectroscopy (LC-MS) has prevailed as the preferred technique for generating high quality metabolomic data. However, even with advanced instrumentation and established data acquisition protocols, technical errors are still routinely encountered and can pose a significant challenge to unveiling biologically relevant information. In large-scale studies, signal drift and batch effects are how technical errors are most commonly manifested. We developed pseudoDrift, an R package with capabilities for data simulation and outlier detection, and a new training and testing approach that is implemented to capture and to optionally correct for technical errors in LC–MS metabolomic data. Using data simulation, we demonstrate here that our approach performs equally as well as existing methods and offers increased flexibility to the researcher. As part of our study, we generated a targeted LC–MS dataset that profiled 33 phenolic compounds from seedling stem tissue in 602 genetically diverse non-transgenic maize inbred lines. This dataset provides a unique opportunity to investigate the dynamics of specialized metabolism in plants.

3.2 Introduction

Metabolomics concerns the study of small molecular compounds or metabolites (<1500 Da) and is essential for advancements in metabolism research in biological systems (Roberts et al., 2012). The metabolome is influenced by complex interactions between the genome, proteome, and transcriptome with the environment (Yang et al., 2021), and evoking changes to the metabolome and the associated effects on metabolic pathways has direct applications to pharmacology, drug development, and improvements to plant productivity, composition, and resilience (Kumar et al., 2017; Manzoni et al., 2016). Maize is among the most productive and economically important crops, with an annual production of over 700 million metric tons globally (Ranum et al., 2014). Many tools and technologies have facilitated previous maize improvement efforts, and future improvements will benefit from metabolomics-based research (Medeiros et al., 2021). In particular, valuable insights are likely to come from larger scale metabolomic experiments and evaluations.

A metabolomic analysis involves identifying and/or quantifying many metabolites simultaneously either by targeted (metabolite identity known) or untargeted (metabolite identity unknown) methods (Roberts et al., 2012). Liquid chromatography coupled with tandem mass spectrometry (LC–MS) is currently the dominant technique used in biological research due to its high sensitivity and selectivity (Sánchez-Illana et al., 2018; Wehrens et al., 2016). However, metabolic profiles from extracts obtained from genetically identical individuals, or even from the same biological sample can have significant variability, the sources of which can be traced to extraction efficiency, changes in the injection volume, inlet interference contamination, column contamination (due to the complex matrices), or drift in ionization efficiency (Wehrens et al., 2016). Despite taking appropriate actions during the data acquisition phase to minimize unwanted sources of variation, obtaining repeatable measurements across different instruments, and even on the same instrument, remains a challenging task. Furthermore, these challenges can be compounded in large-scale studies where LC–MS runs are often split into several analytical batches, each potentially containing a unique combination of technical variability (Kuligowski et al., 2015).

Combining data across batches in large-scale studies is imperative for increasing the statistical power of downstream analyses and interpretation. Therefore, several strategies have emerged to capture and correct for systematic errors in metabolomics data (Han & Li, 2020). The most robust methods rely on regularly interspersed quality control (QC) samples included within and between batches (Broadhurst et al., 2018), which are meant to capture temporal signal drift trends and detect any additional technical errors. QC data points can then be used to apply a computationally efficient correction to the entire dataset, for instance, with a QC-Robust Spline Correction (QC-RSC) (Kirwan et al., 2013). However, there is no universally applicable approach to predetermine QC sample composition or the frequency at which QC samples should be included along the LC–MS run sequence; these are typically influenced by practical and experiment-specific considerations. While a high frequency of QC sample inclusion may be useful for thoroughly capturing systematic errors, these samples occupy LC–MS run slots that might otherwise be allocated to experimental samples, increasing the size of the experiment and subsequently increasing the opportunity for error. To this end, non-QC-based correction approaches have been attempted (Bararpour et al., 2021; Rusilowicz et al., 2016; Wehrens et al., 2016). However, the benefits from increased experimental throughput come at the expense of a potentially reduced ability to detect and correct for systematic errors.

Here, we present pseudoDrift, a new R package that combines the beneficial aspects of QC- and non-QC-based technical error correction methods. pseudoDrift relies on a training and testing procedure to estimate non-observed QC data points (pseudoQC) from metabolomics data with partial QC (trueQC) representation. pseudoDrift is not a correction method per se, but rather offers a strategy for estimating pseudoQC reference points that can be used to make data corrections using other already existing methods. By default, pseudoDrift uses the QC-RSC correction method (Kirwan et al., 2013). However, the non-adjusted data containing pseudoQC estimations are also returned and can be exported if alternative correction methods are preferred by the user. Additional functionality of pseudoDrift comes in the form of simulation and outlier detection capabilities. Compared to currently available LC–MS simulation tools (Bielow et al., 2011; Kösters et al., 2021; Noyce et al., 2013; Schulz-Trieglaff et al., 2008), which aim to simulate raw data that need to subsequently be processed before generating manipulatable peak matrices, pseudoDrift allows for direct simulation of peak matrices based on metadata gathered from online databases such as the MassBank of North America (MoNA, 2022). The outlier detection method implemented in pseudoDrift relies on absolute differences between sample replicates to generate an expected tailed distribution, which is used to identify potential outliers. The pseudoDrift package is publicly available at https://github.com/jrod55/pseudoDrift (accessed on 11 April 2022). In this study, the main objectives were to demonstrate the utility of pseudoDrift using simulated data and to apply the workflow to a newly generated large-scale maize seedling LC–MS phenolic profiling dataset. Although the metabolites extracted and analyzed were from maize, the development and assessment of this tool are applicable to any specific organism. We selected maize because of its importance as a crop and the availability of a well-studied diversity panel (Hansey et al., 2011; Mazaheri et al., 2019).

3.3 Results and Discussion

Simulating data with pseudoDrift

To compare each of the analysis workflow functions from pseudoDrift with existing approaches, we used data produced with the ‘simulate_data()’ function, assuming three batches with sample sizes of 100, 200, and 300, respectively (Figure S1 in Supplemental File 3.1). We set the QC frequency to have each 25th sample represent a trueQC, and used the default effect severity settings (slope and batch magnitude set to 1.25). The simulated data corresponded to the phenolic compound tricin, which we queried using the MoNA (2022) accession ID FIO00738. From the simulated data output, we retained the type 4 signal drift effect data for analysis, and as a control, the data without any signal drift effects (see Materials and Methods). As a reference, and to depict the inputs and outputs of ‘simulate_data()’, we included a visual representation of a simple example of three batches that we did not use in our analysis (Figure 3.1).

Performance evaluation of the pseudoDrift analysis workflow

For outlier detection, we compared ‘pw_outlier()’ to the commonly used 1.5*Interquartile Range (IQR) method, and a formal statistical testing approach with an iterative evaluation of the Grubbs test statistic (Grubbs, 1950) with a 0.05 type 1 error rate. With `pw_outlier`, we used the default 0.95 quantile threshold (Figure 3.2), and we observed the expected tailed distribution in each batch of the simulated data (Figure S2 in Supplemental File 3.1).

Between outlier detection methods tested, ‘pw_outlier()’ identified a total of 24 potential outliers, IQR detected 7, and the iterative Grubbs test identified a single point (Figure S3 in Supplemental File 3.1). Although ‘pw_outlier()’ detected the largest number of possible outliers, they were not all necessarily at the extremes of the peak area distribution. The distinguishing feature of ‘pw_outlier()’, compared to the two conventional outlier detection methods tested, was that it accounted for the experimental unit (set of three observations) as opposed to treating the distribution of all observations as a whole. This was particularly relevant given that extreme observations, when consistent between replicate measures, could represent biological anomalies that may warrant further investigation. Treating all samples as a whole and basing outlier detection on their distribution would not necessarily have captured the within experimental unit variability as ‘pw_outlier()’ did. A noteworthy characteristic of ‘pw_outlier()’ is that the number of potential outliers returned will always be a function of the quantile threshold set. In our simulated data, the 24 potential outliers represented 5% (1-quantile threshold) of non-QC observations from each batch. Therefore, the user should take this into consideration when applying ‘pw_outlier()’ and interpreting the results.

We assessed the performance of ‘pseudo_sdc()’ by using the simulated data batch with the largest sample size as the training batch and then comparing the coefficient of determination (R^2) and root mean squared error (RMSE) from 10-fold cross-validation between signal drift corrected data and the control. Here, the comparison we made was between using either pseudoQC or trueQC samples to correct for signal drift with the QC-RSC (Kirwan et al., 2013) method. As a reference, we included a visual representation of parameters optimized by ‘pseudo_sdc()’ and an example of how the process was carried out (Figure 3.3).

Gauging the performance of ‘pseudo_sdc()’ solely on its ability to reduce the variability among the simulated trueQC samples would have underestimated the ability of pseudoQC samples to capture and correct for signal drift (Figure S4 in Supplemental File 3.1). However, the benefit of using simulated data was that each observation effectively served as a QC sample since we knew the originally simulated value. Regressing the signal drift corrected data on the originally simulated values, we found that using pseudoQCs to correct the data resulted in a correction ($R^2 = 0.7404$; RMSE = 3810) which was on par with the performance of a trueQC-based correction ($R^2 = 0.7499$; RMSE = 3744) (Figure S5 in Supplemental File 3.1). Our data-driven approach, which took advantage of the data variability while training a model on trueQCs as anchors in the training batch, also reduced systematic bias in how the signal drift correction was applied. The trueQC correction more effectively corrected a portion of batch 3 in the simulated data compared to other batches (Figure S5 in Supplemental File 3.1), whereas the pseudoQC correction appeared to have an equal performance across batches (Figure S5 in Supplemental File 3.1). The bias reduction with the pseudoQC correction was likely due to the differences between how trueQC and pseudoQC samples captured the signal drift trend. The trueQC samples were independent of the data as a whole and were able to capture sharp changes in signal drift, while pseudoQC provided a more general representation of the trend and captured more subtle signal drift patterns with information from the data variability.

Maize LC–MS phenolic data analysis with pseudoDrift

Prior to applying the pseudoDrift analysis workflow to the normalized maize LC–MS dataset consisting of peak areas for 33 phenolic compounds (Supplemental Table 3.1), we identified five compounds (apigenidin, dihydrokaempferol, luteolin 7-*O*-glucoside, syringic acid, and syringol) with limits of detection (LOD) threshold values greater than 25% of all experimental samples across sub-batches. The LOD for each compound varied by sub-batch (Supplemental Table 3.1), and, therefore, so did the proportion of experimental samples above the LOD. To avoid a large proportion of missing data for apigenidin, dihydrokaempferol, luteolin 7-*O*-glucoside, syringic acid, and syringol, they were completely excluded from downstream analyses. The remaining compounds were independently analyzed with the ‘pw_outlier()’

function with default arguments. This identified the top 5% of observations per batch as possible outliers. Importantly, with the conservative action of omitting all observations identified with ‘pw_outlier()’, no single inbred line was removed completely from the data. The ‘pw_outlier()’ cleaned data were subsequently processed with ‘pseudo_sdc()’ with batch 4 used as the training batch. The optimal parameters to estimate pseudoQC samples were determined for each compound (Supplemental Table 3.2) and used to apply the signal drift correction across batches (Supplemental File 3.2). Each compound had distinct signal drift patterns, although the batch-to-batch effect was substantially more pronounced for some compounds. For example, caffeic acid and 4-chlorogenic acid in batch 2 were considerably lower than in other batches (Supplemental File 3.2). However, since ‘pseudo_sdc()’ calculated pseudoQC samples based on quantiles determined in the training batch, we were able to capture this batch-to-batch effect along with the signal drift trends.

With existing QC-based signal drift and batch correction methods such as QC-RSC (Kirwan et al., 2013), we would have been restricted to analyzing the data from batch 4 alone since it was the sole batch with trueQC samples represented (Table 3.1). Compared to the two non-QC-based methods tested, combatting batch effects (ComBat) (Johnson et al., 2007; Leek et al., 2012) and batch effect removal (ber) (Giordan, 2014), pseudoDrift substantially reduced the maximum distance between any two trueQC points when plotted along the first and second principal components (PCs) (Figure 3.4). Thus, suggesting an overall improved correction across compounds.

Independently, even for compounds with severe batch-to-batch effects, such as caffeic acid and 4-chlorogenic acid (Supplemental File 3.2), there were no major differences in the inter-batch corrections between pseudoDrift and the ComBat and ber methods (Figure 3.5). However, as demonstrated by a flatter line among trueQC points in batch 4, pseudoDrift performed best at simultaneously correcting for intra-batch signal drift effects as well. Together, these results highlight the improvements to signal drift and batch corrections, which pseudoDrift achieved by coalescing QC and non-QC approaches into a new correction method.

3.4 Materials and Methods

Plant material and experimental design

We grew a set of 602 genetically diverse maize inbred lines in a controlled environment room under high-intensity light emitting diode (LED) lights at the Wisconsin Crop Innovation Center. Our experiment consisted of a completely randomized design (CRD), with three replications per inbred line, with independent randomization applied per set of 602 lines. We recorded instances where seeds failed to germinate in any of the first three replicates and re-planted these in a fourth replication. In all replications, we sowed seeds into 32-cell flats and hand watered them every other day for the first 7 days. On the 8th day, we transitioned to a daily watering with an automated flood fertigation watering system, which was programmed to submerge the 32-cell flats for 5 min per day. A total of 21 days passed from planting to harvest. During the growing period, we maintained the temperature at 28 °C and artificially controlled the photoperiod by supplying 16 h of light followed by 8 h of darkness. At harvest, we used 15 mL conical Falcon tubes to collect the basal 6 cm of seedling stem tissue and immediately placed samples into liquid nitrogen prior to lyophilizing.

Reagents for stock and working solutions

We procured the reagents used in this study 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), and when performing dilutions, we used ultrapure (>18 Ω) water generated through a Milli-Q system. The metabolites we profiled included 33 phenolic compounds with available chemical standards (Supplemental Table 3.1). For each compound, we prepared 1 mM stock solutions by reconstituting them in 80% (*v*/*v*) HPLC grade methanol or 100% dimethyl sulfoxide (DMSO). We prepared a pooled mixture containing all 33 standards, each at 1 μM, and through serial dilution produced samples with concentrations between 1000 nM and 1.7 nM that we used as external standards. Following the same procedure, but with a final concentration of 50 nM, we prepared an internal standard (8-prenylnaringenin). For sample preparation, we used an extraction solvent consisting of 80% (*v*/*v*) HPLC grade methanol and 0.1% (*v*/*v*) formic acid.

Preparation of stem tissue extracts and QC samples

The sample preparation occurred in four separate batches, each including a different number of experimental samples (Table 3.1). Batch 1 was the smallest with 165 samples, followed by batch 2 with 198 samples, batch 3 with 663 samples, and batch 4 with 1008 samples. We homogenized the maize seedling stems using liquid nitrogen and PVC tubes containing a metal bead which a paint shaker (5G-HD Harbil 5-Gallon Shaker model 37600) agitated for 2 min at 60 Hz. To avoid cross-contamination, we washed the PVC tubes and metal beads with distilled water and soap between samples. We then transferred a ~50 mg subset of the homogenized plant material to a 2 mL Eppendorf tube and combined it with the extraction solvent. Batch 4 included the 8-prenylnaringenin internal standard, which was added at the same time that we combined the extraction solvent and the plant material. Our extraction protocol consisted of a 12 h incubation at 4 °C, followed by reconstitution by vortexing for 20 s, centrifugation for 5 min at 15,000× *g* at room temperature, and recovery of the supernatant for analysis by LC–MS. We prepared a QC sample from 100 randomly selected samples and included it in the analysis for batch 4 alone.

LC–MS data acquisition

The instrument used for data acquisition was a Waters ACQUITY TQD Tandem Quadrupole UPLC/MS/MS (Waters Corporation, Milford, MA, USA). We created a 10 min targeted multiple reaction monitoring (MRM) method for detecting 33 phenolic compounds (Supplemental Table 3.1) and ran samples in accordance with their corresponding preparation batch. The method was based on a modification of the MRM method previously described (Cocuron et al., 2019). Within batches, we designated samples to sub-batches based on the preceding external standards set (Figure 3.6). Across the four batches, we ran a total of 13 external standards sets. For each inbred maize line, we ran the three biological replicates consecutively of one another (e.g., Line 1 rep1, then Line 1 rep2, Line 1 rep 3, Line 2 rep1, etc.). While in the queue, samples and external standards remained at −10 °C in an autosampler. The instrument used a 10 µL injection volume, and the liquid chromatographic separation occurred at 30 °C using a reverse phase Waters Symmetry C18 column (4.6 × 75 mm; 3 µm) with a Symmetry C18 prep-column (3.9 × 20 mm; 5 µm) (Waters Corporation, Milford, MA, USA). We integrated peak areas for all compounds using MassLynx (v 4.2) with vendor-specific data files (.raw) to produce the raw peak area matrix.

PseudoDrift workflow

The pseudoDrift R package consisted of three main functions, including ‘simulate_data()’, ‘pw_outlier()’, and ‘pseudo_sdc()’. We wrote these functions to run independently of one another, although here we defined the analysis workflow as applying the ‘pw_outlier()’, and ‘pseudo_sdc()’ functions in sequential order.

We wrote the ‘simulate_data()’ function to accept a structure-data file (SDF) as input, such as those obtained from MoNA (2022), and to return the queried compound metadata as output, along with a simulated peak area matrix, and four distinct signal drift and batch effect types applied to the simulated data. The simulated peak area patterns among QC samples were used to define the four effect types. A monotonic increase or decrease was characteristic of a type 1 effect, a type 2 effect described changes in magnitude occurring between batches, a type 3 effect was random, and a type 4 effect consisted of a combination of type 1 and type 2 effects and represented what is most commonly encountered in metabolomics datasets. To provide user flexibility and ensure reproducible simulation results, we included arguments for seed setting, QC frequency, batch size, and effect type severity.

The first analysis function we developed was ‘pw_outlier()’, which served as an outlier detection method to accommodate common features of metabolomic data, including skewed distributions and limited biological or technical replication. Our approach relied on assessing pairwise absolute differences within and between replicate measures of samples. To illustrate, we considered a hypothetical metabolite or feature and a single sample with three biological replicates. The ‘pw_outlier()’ function computed the pairwise differences between replicates as |rep1-rep2|, |rep1-rep3|, and |rep2-rep3| and then extended this computation to each sample within a given batch. This generated a distribution of pairwise differences, which we assumed to be positively skewed with observations at the upper tail representing potential outliers. The default threshold, which we set for `pw_outlier` to use, was the 0.95 quantile of all sample pairwise differences in a batch. We included arguments allowing the user to have flexibility over the quantile threshold and grouping factor used for calculations.

We developed the second analysis function, ‘pseudo_sdc()’, to work with a variable representation of trueQC samples across batches. Using data corresponding to the batch where trueQC samples were most represented, or the batch designated as the training batch, ‘pseudo_sdc()’ optimized four parameters that were used to calculate pseudoQC points that effectively captured the signal drift pattern in the training batch. To effectively capture signal drift in this context meant to minimize the criterion set by the user, which, by default, we set to use the mean squared error (MSE) between the trueQC samples and the estimated pseudoQC points in the training batch. The four parameters, which ‘pseudo_sdc()’ optimized, concerned the range of values used in the calculation (quantile.increment), the number of equally sized breaks to divide the batch into (test.breaks), the window size to calculate a rolling median over (test.window), and the positional offset (test.index) of pseudoQCs, relative to test.window. With the optimized parameters from the training batch, ‘pseudo_sdc()’ applied the same parameters to the remaining batches to estimate the pseudoQC points from the data variability. We integrated the QC-RSC (Kirwan et al., 2013) method into ‘pseudo_sdc()’ with an auxiliary function from the pmp R package (Jankevics et al., 2022) to perform the data correction with pseudoQC samples in lieu of the trueQC samples.

To illustrate the functionality of pseudoDrift, we included a tutorial (Supplemental File 3.1), where we walked through each of the analysis functions of pseudoDrift. In the tutorial, we used several additional R packages, including ChemmineR (Cao et al., 2022) for SDF file indexing, ggpubr (Kassambara, 2020) and cowplot (Wilke, 2020) for plotting, data.table (Dowle et al., 2021) and tidyverse (Wickham et al., 2019) for data manipulation, and caret (Kuhn et al., 2022) for regression modeling.

LC–MS data normalization and processing with pseudoDrift

We set the limits of detection (LOD) for each phenolic compound as three times the peak area of the blank (extraction solvent alone) and established the thresholds on a per compound and sub-batch basis, with the blank reference value for each compound calculated as the mean peak area across blank samples within the corresponding sub-batch. If cumulatively across sub-batches more than 25% of samples were below the respective compound LOD thresholds, we removed the compounds from the data matrix and completely excluded them from the downstream analyses. Rather than using absolute values, we opted for relative peak area values to ensure compounds were uniformly analyzed, including those that accumulated at high levels in maize stem tissues (outside the upper range of external standards). We normalized the data by the weight of each sample to provide arbitrary units of area (AUA) and removed blanks and external standards from the data matrix prior to analyzing with the pseudoDrift workflow. When applying the analysis workflow, we processed one compound at a time, first with ‘pw_outlier()’, then with ‘pseudo_sdc()’ using batch 4 as the training batch to estimate pseudoQC samples across all batches, and to correct for signal drift and batch effects in the data. We tested two additional non-QC correction methods, specifically the ComBat (Johnson et al., 2007; Leek et al., 2012) and ber (Giordan, 2014) methods implemented in the dbnorm R package (Bararpour et al., 2021), and compared the data corrections based on the maximum distance of trueQC points (maxDist) along the first two PCs calculated from the full peak area matrix. A smaller maxDist indicated trueQC samples had less variability, and thus, a better correction of technical errors.

3.5 Conclusions

The number and impact of MS-based metabolomics studies in the biological sciences are likely to rise as methods improve and accessibility to instrumentation by researchers increases. This is particularly true for systems biology, which now has a plethora of complementary omics tools available to investigate previously unexplored areas of research. In metabolomics, however, there are still currently various limitations, which, if not addressed, result in abnormally noisy data. Here, we developed a simulation and analysis tool for applying statistical techniques in a training and testing framework, to calculate and correct for technical errors in a dataset and to identify potential outliers. We applied this analysis tool to maize phenolic compounds, including phenylpropanoids and flavonoids, since they play important functions in the interaction of maize with the environment and provide health benefits to humans (Dong & Lin, 2021; Jiang et al., 2016; Parihar et al., 2015). The here-developed tool has numerous applications, such as combining datasets across studies with differing levels of trueQC sample representation, identifying irregular observations in data, and as an experiment planning resource. An advantage of pseudoDrift is that it is written in R and includes an extensive tutorial (Supplemental File 3.1), which makes it accessible to all users, including those without extensive programming experience. Since pseudoDrift uses a train–test procedure, a disadvantage might come from users attempting to apply the method to estimate pseudoQC points from small training batches. To offer the greatest flexibility to users, pseudoDrift does not have any batch restrictions, but rather we include a warning to users within the software documentation in R. While the focus of our study was on a targeted LC–MS method applied to samples prepared from a very large number of maize seedlings, the methods described can be broadly applied to other metabolomics datasets, or any temporally variable data prone to technical errors.

3.6 Tables

Table 3.1: Summary of samples per batch, type of standard included, and whether QC samples were represented or not.

Batch	**Num. Samples**	**External Standard**	**Internal Standard**	**QC samples represented**
B1	165	Yes	No	No
B2	198	Yes	No	No
B3	663	Yes	No	No
B4	1008	Yes	Yes	Yes

3.7 Figures

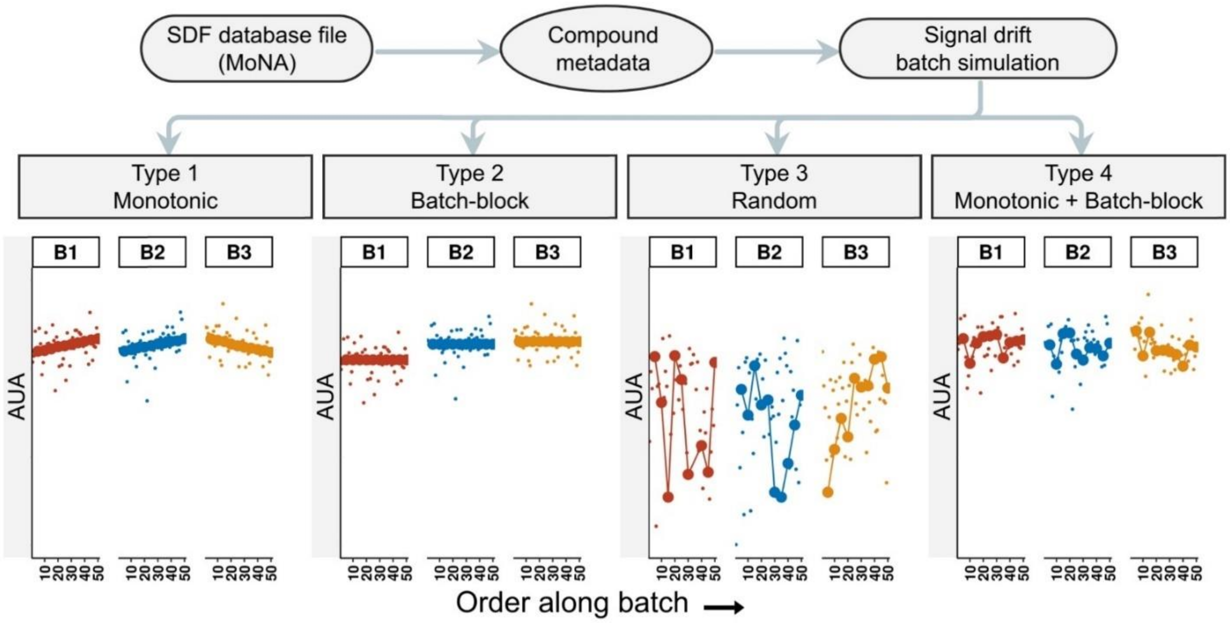


Figure 3.1: Simulated data produced by the ‘simulate_data()’ function. For each compound queried in the user provided structure-data file (SDF), a simulated peak area matrix with arbitrary units of area (AUA) is returned, along with four additional matrices, each with a different signal drift effect type applied. SDF files are available through the MassBank of North America (MoNA). The larger points in each plot represent the simulated QC samples, and smaller points represent non-QC simulated data points.

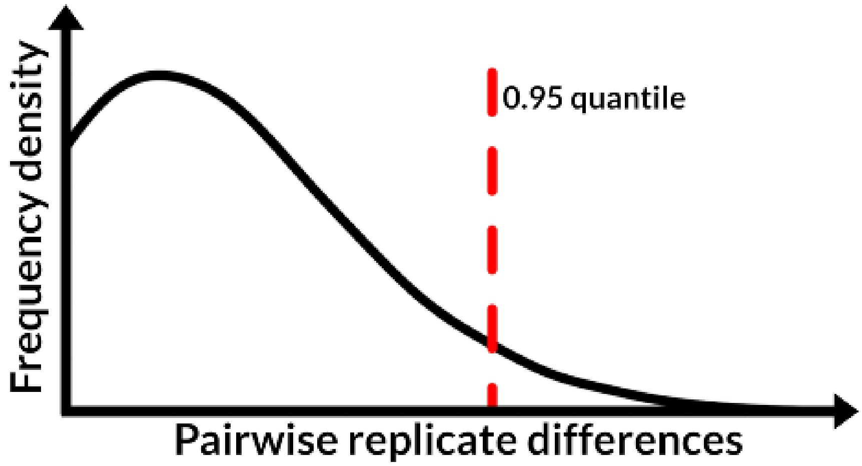


Figure 3.2: Visual representation of the outlier detection approach implemented in the 'pw_outlier()' function. Given a peak area matrix for a particular compound or feature, all pairwise differences between sample replicates are computed. The distribution of these differences is expected to be positively skewed, with values surpassing a given quantile threshold (0.95% shown) marked as potential outliers.

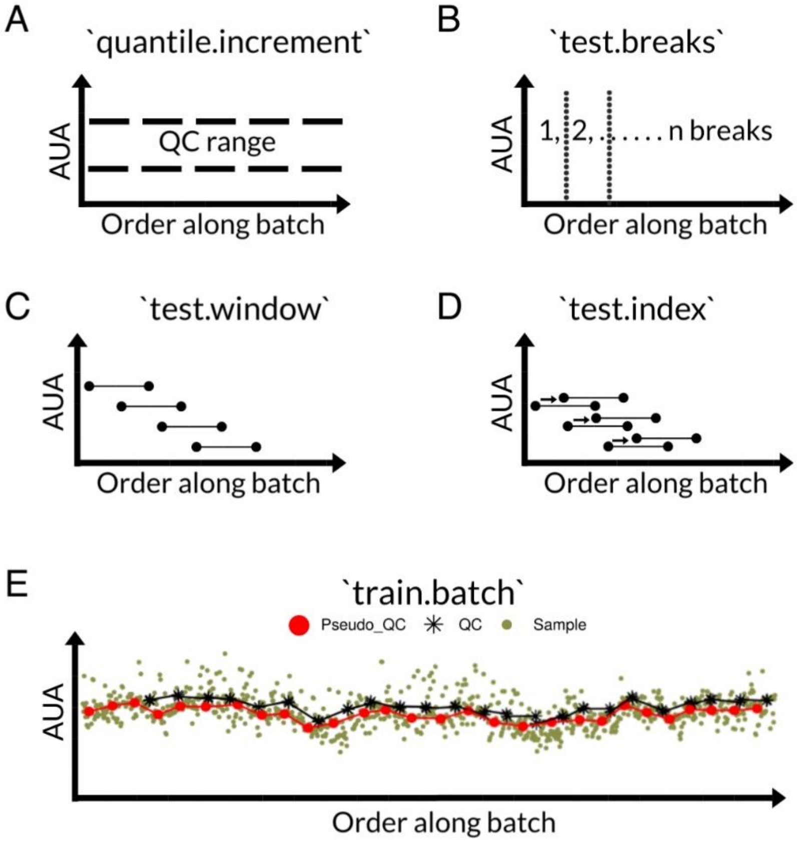


Figure 3.3: The parameters (A–D) optimized for ‘pseudo_sdc()’ that minimize the criteria set by the user. The parameter name is given in the title of each plot, and all computations are made using arbitrary units of area (AUA). By default, the mean squared error is minimized between estimated pseudoQC and true QC samples in the training batch (E).

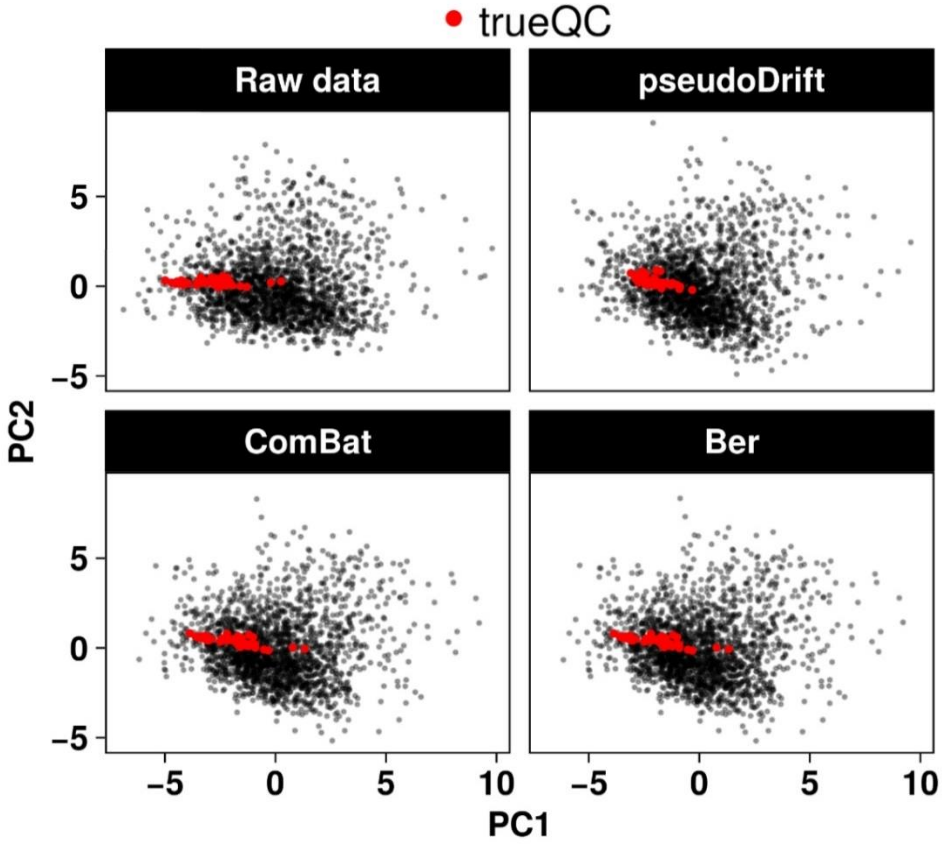


Figure 3.4: Comparison of different non-QC-based corrections applied to the maize LC–MS peak area matrix. The correction method and resultant peak area matrix used to calculate the first two PCs is labeled on each respective bi-plot. The maximum distance (maxDist) represents the largest distance between any two trueQC points in the bi-plot.

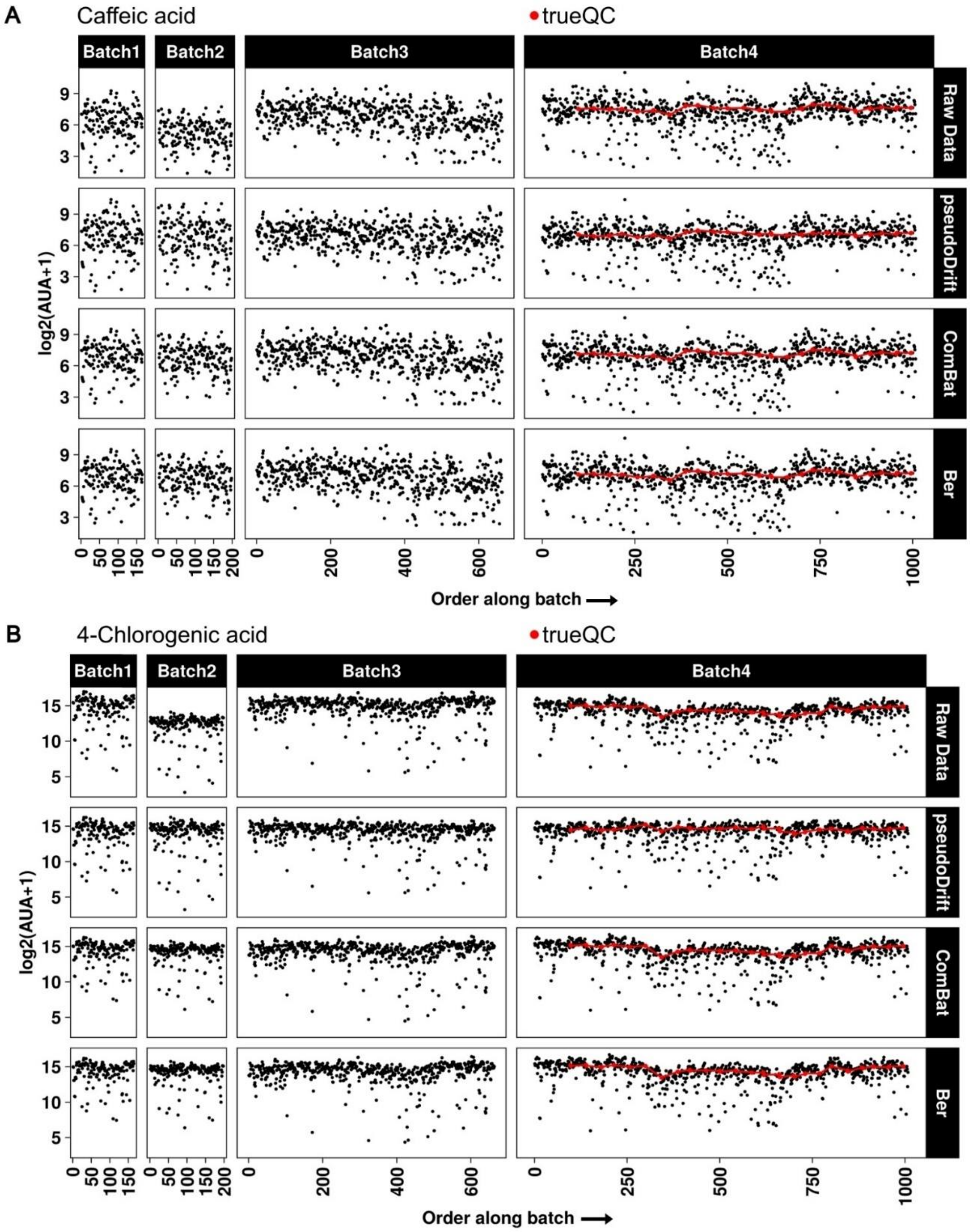


Figure 3.5. Batch-to-batch and intra-batch signal drift effects before and after correction with different non-QC-based corrections. Plotted are the log2 transformed arbitrary units of area (AUA) for caffeic acid (A) and 4-chlorogenic acid (B) before correction (raw data) and after correction with each labeled approach.

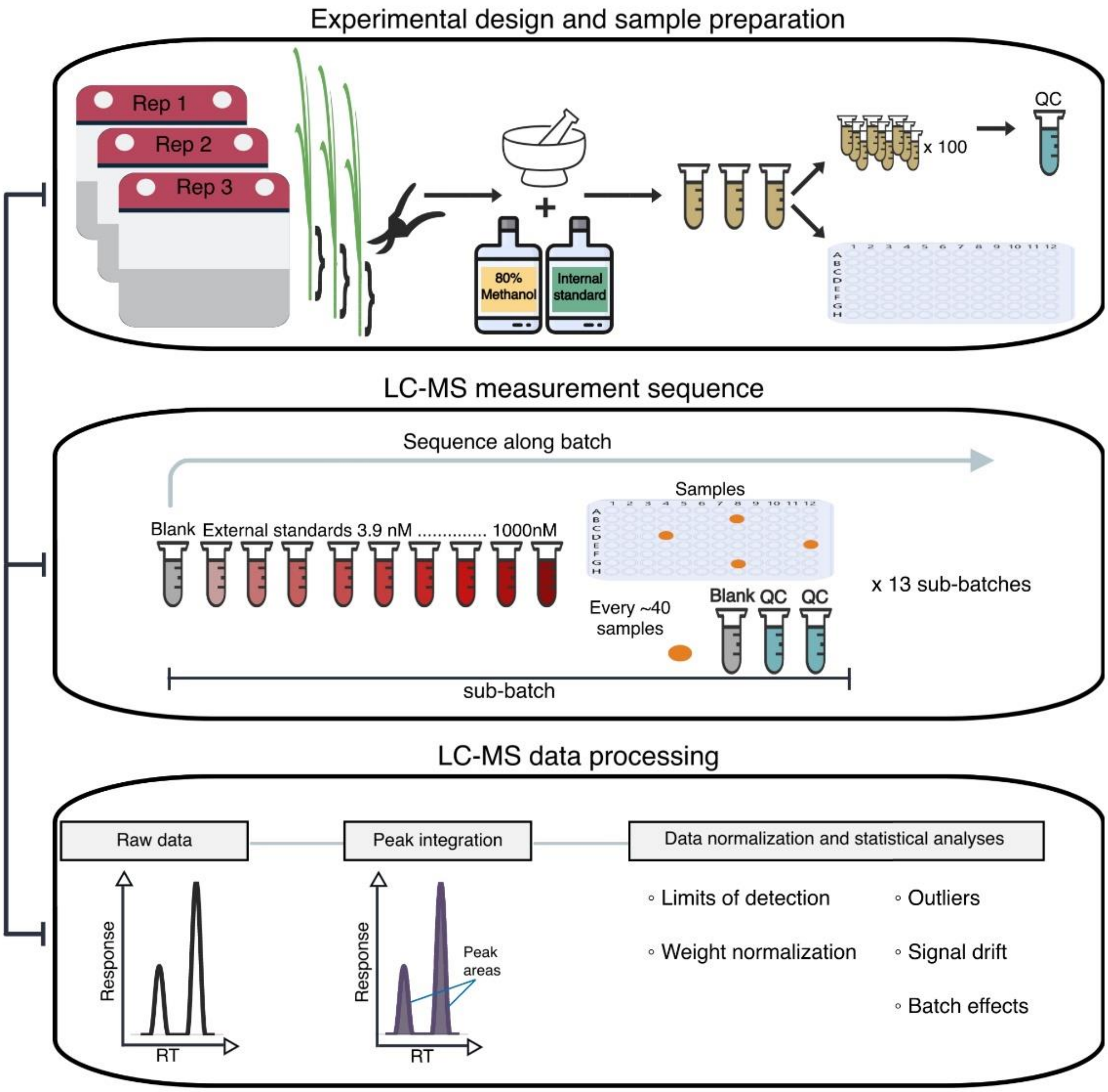


Figure 3.6. Experimental design, data acquisition and data processing steps. Samples labeled as QC represent a pooled mixture from 100 randomly selected experimental samples. A representation of the LC-MS measurement sequence for each batch and sub-batch during the data acquisition phase is shown. During data processing, the depiction represents the instrument response as a function of the retention time (RT).

3.8 Supplementary Materials

The following supplemental files and tables can be downloaded at:

https://www.mdpi.com/article/10.3390/metabo12050435/s1

Supplemental File 3.1: Tutorial for pseudoDrift and accompanying supplementary figures corresponding to simulated data and analyses conducted with pseudoDrift functions.

Supplemental File 3.2: Signal drift trends for all compounds retained after data cleaning. In each plot, arbitrary units of area (AUA) are plotted against the injection order. On each page, (A) is the data before applying the signal drift and batch correction with pseudoQCs, and (B) is the plot after applying the correction.

Supplemental Table 3.1: Targeted multiple reaction monitoring (MRM) method for 33 phenolic compounds investigated in maize seedling (21-day old) stem tissue. Included for each compound are the name, CAS Registry Number, retention time, polarity mode used during data acquisition, MS/MS (+/-) fragments, and the molecular weight. Additional columns designate the limit of detection (LOD) determined for each sub-batch.

Supplemental Table 3.2: Parameters solved for ‘pseudo_sdc()’ when applied to the maize seedling (21-day-old) data, which minimize the mean squared error between pseudoQC and trueQC samples when using batch 4 (B4) as the training batch.

3.9 Author Contributions

Conceptualization, E.G. and N.dL.; methodology, J.R. and L.G.-C.; software, J.R.; validation, J.R.; formal analysis, J.R.; investigation, J.R. and L.G.-C.; resources, E.G. and N.dL.; data curation, J.R.; writing—original draft preparation, J.R.; writing—review and editing, J.R., L.G.-C., E.G. and N.dL.; visualization, J.R.; supervision, E.G. and N.dL.; project administration, E.G. and N.dL.; funding acquisition, E.G. and N.dL. All authors have read and agreed to the published version of the manuscript.

3.10 Funding

This work was supported by the National Science Foundation under Grant No. 1733633. J.R. was supported by a University of Wisconsin-Madison SciMed GRS fellowship. Support for this fellowship is provided by the Graduate School, part of the Office of Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison, with funding from the Wisconsin Alumni Research Foundation. L.G.-C. was supported by a fellowship from Michigan State University under the Training Program in Plant Biotechnology for Health and Sustainability (T32-GM110523).

3.11 Data Availability

The data The data underlying this study, including .raw data files, intermediate and final peak area matrices are openly available through the CyVerse Data Commons at https://doi.org/10.25739/e1bq-kh07 (accessed on 11 May 2022). The code used for normalizing and applying the pseudoDrift workflow to the maize phenolic LC–MS data is available at https://github.com/jrod55/m_lcms (accessed on 11 May 2022).

3.12 Acknowledgments

We thank the Michigan State University Mass Spectrometry and Metabolomics Core Facility for technical assistance in the LC-MS data acquisition. Additionally, we thank Alexander Silva and Summer Blanco for their assistance with sample preparation and processing.

3.13 References

Bararpour, N., Gilardi, F., Carmeli, C., Sidibe, J., Ivanisevic, J., Caputo, T., Augsburger, M., Grabherr, S., Desvergne, B., Guex, N., Bochud, M., & Thomas, A. (2021). DBnorm as an R package for the comparison and selection of appropriate statistical methods for batch effect correction in metabolomic studies. *Scientific Reports, 11*(1), 1–13. https://doi.org/10.1038/s41598-021-84824-3

Bielow, C., Aiche, S., Andreotti, S., & Reinert, K. (2011). MSSimulator: Simulation of mass spectrometry data. *Journal of Proteome Research*, *10*(7), 2922–2929. https://doi.org/10.1021/pr200155f

Broadhurst, D., Goodacre, R., Reinke, S. N., Kuligowski, J., Wilson, I. D., Lewis, M. R., & Dunn, W. B. (2018). Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, *14*(6), 72. https://doi.org/10.1007/s11306-018-1367-3

Cao, Y. E., Horan, K., Backman, T., & Girke, T. (2022). *ChemmineR: Cheminformatics Toolkit for R*. https://doi.org/10.18129/B9.bioc.ChemmineR

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

Dong, N.-Q., & Lin, H.-X. (2021). Contribution of phenylpropanoid metabolism to plant development and plant–environment interactions. *Journal of Integrative Plant Biology*, *63*(1), 180–209. https://doi.org/10.1111/jipb.13054

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., Iwasaki, W., Wenchel, S., … Schwen, B. (2021, September 27). *data.table: Extension of "data.frame."* https://CRAN.R-project.org/package=data.table

Giordan, M. (2014). A Two-Stage Procedure for the Removal of Batch Effects in Microarray Studies. *Statistics in Biosciences*, *6*(1), 73–84. https://doi.org/10.1007/s12561-013-9081-1

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, *21*(1), 27–58. https://doi.org/10.1214/aoms/1177729885

Han, W., & Li, L. (2020). Evaluating and minimizing batch effects in metabolomics. *Mass Spectrometry Reviews*, 1–22. https://doi.org/10.1002/mas.21672

Hansey, C. N., Johnson, J. M., Sekhon, R. S., Kaeppler, S. M., & de Leon, N. (2011). Genetic diversity of a maize association population with restricted phenology. *Crop Science*, *51*(2), 704–715. https://doi.org/10.2135/cropsci2010.03.0178

Jankevics, A., Lloyd, G. R., & Weber, R. J. M. (2022). *pmp: Peak Matrix Processing and signal batch correction for metabolomics datasets*. https://doi.org/10.18129/B9.bioc.pmp

Jiang, N., Doseff, A. I., & Grotewold, E. (2016). Flavones: From Biosynthesis to Health Benefits. *Plants (Basel, Switzerland)*, *5*(2), E27. https://doi.org/10.3390/plants5020027

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kassambara, A. (2020, June 27). *ggpubr: "ggplot2" Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr

Kirwan, J. A., Broadhurst, D. I., Davidson, R. L., & Viant, M. R. (2013). Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. *Analytical and Bioanalytical Chemistry*, *405*(15), 5147–5157. https://doi.org/10.1007/s00216-013-6856-7

Kösters, M., Leufken, J., & Leidel, S. A. (2021). SMITER-A Python Library for the Simulation of LC-MS/MS Experiments. *Genes*, *12*(3), 396. https://doi.org/10.3390/genes12030396

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2022, March 11). *caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret

Kuligowski, J., Sánchez-Illana, Á., Sanjuán-Herráez, D., Vento, M., & Quintás, G. (2015). Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC). *Analyst*, *140*(22), 7810–7817. https://doi.org/10.1039/C5AN01638J

Kumar, R., Bohra, A., Pandey, A. K., Pandey, M. K., & Kumar, A. (2017). Metabolomics for Plant Improvement: Status and Prospects. *Frontiers in Plant Science*, *8*, 1302. https://doi.org/10.3389/fpls.2017.01302

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, *28*(6), 882–883. https://doi.org/10.1093/bioinformatics/bts034

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. (2016). Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, *19*(2), 286–302. https://doi.org/10.1093/bib/bbw114

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., Robin Buell, C., de Leon, N., & Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, *19*(1), 1–17. https://doi.org/10.1186/s12870-019-1653-x

Medeiros, D. B., Brotman, Y., & Fernie, A. R. (2021). The utility of metabolomics as a tool to inform maize biology. *Plant Communications*, *2*(4), 100187. https://doi.org/10.1016/j.xplc.2021.100187

MoNA. (2022, May 12). https://mona.fiehnlab.ucdavis.edu/

Noyce, A. B., Smith, R., Dalgleish, J., Taylor, R. M., Erb, K. C., Okuda, N., & Prince, J. T. (2013). Mspire-Simulator: LC-MS shotgun proteomic simulator for creating realistic gold standard data. *Journal of Proteome Research*, *12*(12), 5742–5749. https://doi.org/10.1021/pr400727e

Parihar, A., Grotewold, E., & Doseff, A. I. (2015). Flavonoid Dietetics: Mechanisms and Emerging Roles of Plant Nutraceuticals. In C. Chen (Ed.), *Pigments in Fruits and Vegetables: Genomics and Dietetics* (pp. 93–126). Springer. https://doi.org/10.1007/978-1-4939-2356-4_5

Ranum, P., Peña-Rosas, J. P., & Garcia-Casal, M. N. (2014). Global maize production, utilization, and consumption. *Annals of the New York Academy of Sciences*, *1312*, 105–112. https://doi.org/10.1111/nyas.12396

Roberts, L. D., Souza, A. L., Gerszten, R. E., & Clish, C. B. (2012). Targeted Metabolomics. *Current Protocols in Molecular Biology*, *98*(1), 30.2.1-30.2.24. https://doi.org/10.1002/0471142727.mb3002s98

Rusilowicz, M., Dickinson, M., Charlton, A., O’Keefe, S., & Wilson, J. (2016). A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. *Metabolomics*, *12*(3), 56. https://doi.org/10.1007/s11306-016-0972-2

Sánchez-Illana, Á., Piñeiro-Ramos, J. D., Sanjuan-Herráez, J. D., Vento, M., Quintás, G., & Kuligowski, J. (2018). Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling. *Analytica Chimica Acta*, *1019*, 38–48. https://doi.org/10.1016/j.aca.2018.02.053

Schulz-Trieglaff, O., Pfeifer, N., Gröpl, C., Kohlbacher, O., & Reinert, K. (2008). LC-MSsim – a simulation software for liquid chromatography mass spectrometry data. *BMC Bioinformatics*, *9*(1), 423. https://doi.org/10.1186/1471-2105-9-423

Wehrens, R., Hageman, Jos. A., van Eeuwijk, F., Kooke, R., Flood, P. J., Wijnker, E., Keurentjes, J. J. B., Lommen, A., van Eekelen, H. D. L. M., Hall, R. D., Mumm, R., & de Vos, R. C. H. (2016). Improved batch correction in untargeted MS-based metabolomics. *Metabolomics*, *12*, 88. https://doi.org/10.1007/s11306-016-1015-8

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C. O. (2020, December 30). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2."* https://CRAN.R-project.org/package=cowplot

Yang, Y., Saand, M. A., Huang, L., Abdelaal, W. B., Zhang, J., Wu, Y., Li, J., Sirohi, M. H., & Wang, F. (2021). Applications of Multi-Omics Technologies for Crop Improvement. *Frontiers in Plant Science*, *12*, 1846. https://doi.org/10.3389/fpls.2021.563953

Chapter 4: Genetic architecture of lignin precursors and related phenolic metabolites in diverse maize lines

Authors:

Jonas Rodriguez[1], Lina Gomez-Cano[2], Erich Grotewold[2], Natalia de Leon[1]

Affiliations:

[1] Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA

[2] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, 48824, USA

4.1 Abstract

Secondary metabolites in plants are foundational for their role in growth and development, defense, and adaptation. Among the most abundant and diverse are phenolic compounds, which comprise a large class of secondary metabolites including phenylpropanoids, flavonoids, and lignins. These compounds are important in defining agronomic, biomass compositional, and nutritional traits. This study utilized genetic and transcriptomic analysis to dissect quantitative variation in four phenolic compounds known to integrate into the lignin polymer of maize specifically coniferyl aldehyde, ferulic acid, sinapic acid, and tricin. We coupled genome-wide and transcriptome-wide association studies along with differential gene expression analysis to identify and prioritize candidate genes associated with phenolic accumulation. The data analyzed was generated by liquid chromatography-mass spectroscopy from stem tissues of 21-day old seedlings of 597 diverse inbred lines and four transgenic RNAi lines. Phenolic metabolite repeatability estimates were moderate to high (0.64 - 0.77) among the diverse maize lines, supporting the potential for transcriptomic and genetic dissection. We identified a total of 20 high confidence candidate genes associated with phenolic accumulation. Among these, known phenolic biosynthesis genes like phenylalanine ammonia lyase appeared and served as positive controls in our analysis. Of the genes identified without known functions, multiple sources of evidence linked a zinc finger protein gene to quantitative variation for sinapic acid accumulation. The list of high confidence candidate genes identified in our study serves as a valuable resource to guide experimental work aiming to validate aspects of phenolic compound biosynthesis that are not entirely known such as pathway regulation.

4.2 Introduction

Plants harbor the largest biochemical diversity encountered in nature. The phenylpropanoid pathway in particular is responsible for producing a diverse group of biomolecules which are important for specialized as well as general functional metabolism that modulates plant growth, development, and defense (Deng & Lu, 2017; Vogt, 2010). Lignin is an example of a crucial biopolymer needed in land plants for structural support and water solute transport through the plant vasculature system (Lei, 2017). Biosynthesis of lignin relies on three canonical phenolic compounds (*p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol) known as monolignols, which are biosynthesized from a branch of the phenylpropanoid pathway (Boerjan et al., 2003; Ralph et al., 2004). Recent work in maize has demonstrated that alternative monolignol phenolics such as tricin and ferulic acid can also integrate into the lignin polymer, both natively and as a response to reduced biosynthesis of canonical monolignols (Karlen et al., 2016; Lan et al., 2016; Liu et al., 2021).

There are potential economic, environmental, and human health benefits associated with modifying the phenolic profiles of plants. For example, plant biomass with an enhanced abundance of valuable hydroxycinnamic acids (e.g., *p*-coumaric acid or ferulic acid) can offset operational costs and reduce waste from lignocellulosic biorefineries (Karlen et al., 2020; Kumar et al., 2018). The value of hydroxycinnamic acids is in part derived from their prophylactic and antioxidant properties and their incorporation into some supplements and pharmaceuticals (Alam et al., 2016). Other examples of medical applications that utilize lignin-derived compounds include hydrogels and 3-D printed meshes for treating wound patients (Domínguez-Robles et al., 2019; Laftah et al., 2011).

In maize, the enzymatic genes involved in monolignol biosynthesis have been relatively well studied, primarily by using reverse genetic approaches involving mutant and transgenic plants to reveal gene function (Barrière et al., 2015; Vanholme et al., 2012; Wang et al., 2015). Among the most well characterized mutants in maize are the *brown midrib* (*bm*) mutants, which present a distinct phenotype consisting of a red-brown coloration in the vasculature tissue of the leaf midrib. There are several maize *bm* mutants, and they are historically relevant for their roles in forage and biomass breeding programs (Barrière et al., 1993). The most widely used maize *bm* mutants are the *bm1* and *bm3* mutants, which disrupt the production of cinnamyl alcohol dehydrogenase (CAD) and caffeic acid-3-O-methyltransferase (COMT), respectively. The *bm1* maize mutant, in particular, presents a unique opportunity to study the final steps in monolignol biosynthesis because of the role that CAD plays in the final reduction of hydroxycinnamaldehydes (including coniferyl aldehyde) to monolignol alcohol subunits that make up the majority of the lignin polymer (Mansell et al., 1976; Morrison et al., 1994).

In a recent study, Liu et al. (2021) evaluated a *bm1* mutant inbred at silage maturity and observed a notable increase in ferulic acid, sinapic acid, and tricin. They speculated that their observations of elevated ferulic acid and sinapic acid concentrations might be explained by the increased enzymatic activity of hydroxycinnamaldehyde dehydrogenase (HCALDH), similar to reports in Arabidopsis (Nair et al., 2004) and *Brassica napus* (Mittasch et al., 2013). The enzymes encoded by the four closest Arabidopsis HCALDH homologs in maize have been demonstrated *in vitro* to catalyze coniferyl aldehyde and sinapyl aldehyde to ferulic acid and sinapic acid, respectively (Končitíková et al., 2015). However, it is unknown if increased gene expression or enzymatic activity of HCALDH *in vivo* can increase ferulic acid and sinapic acid concentrations in maize. Liu et al. (2021) offered an alternative explanation for the increase in ferulic acid alone, suggesting that there may be a physical interaction between CAD and the upstream cinnamoyl CoA reductase (CCR) enzyme, where a reduction in CAD is coupled with a reduction in CCR, as has been demonstrated in Poplar (Yan et al., 2019). The reduced expression of both CAD and CCR was speculated to result in the accumulation of feruloyl-CoA, which would then favor the conversion to ferulic acid.

In contrast to the enzymatic genes involved in monolignol biosynthesis, a comprehensive dissection of the regulatory genes that orchestrate the expression of the enzymatic genes is somewhat more complex. For example, using chromatin immunoprecipitation followed by sequencing (ChIP-seq) it has been demonstrated that the transcription factor (TF) LBD35 binds with the gene promoters of 24 enzymatic genes from the phenylpropanoid pathway (F. Yang et al., 2017). The genes involved in monolignol biosynthesis that LBD35 binds to include enzyme encoding genes that are active at the start, middle, and end of the pathway, including phenylalanine ammonia lyase (PAL), CCR, and CAD, respectively (F. Yang et al., 2017). ChIP-seq and other *in vivo* approaches are undoubtedly essential for dissecting gene regulation, and larger ChIP-Seq datasets in maize are now becoming available (Galli et al., 2018; Ricci et al., 2019; Tu et al., 2020; F. Yang et al., 2017). However, necessary validation experiments are still technically challenging and cost-prohibitive to perform for all known TFs in maize. Therefore, identifying promising candidates for validation remains critical.

A promising approach for identifying candidate genes for functional validation includes leveraging existing datasets to perform genome-wide association studies (GWAS) or a combined approach of GWAS with transcriptome-wide association studies (TWAS) using metabolomic data as a molecular phenotype. Here, we characterized the natural variation of four phenolic compounds including coniferyl aldehyde, ferulic acid, sinapic acid, and tricin among diverse inbred maize lines. We then performed GWAS and TWAS for these compounds to identify high confidence candidate genes associated with their accumulation. In addition, we conducted a differential gene expression analysis (DEA) using *bm1* RNAi lines to support the association analyses and to test for differential expression of CCR and HCALDH encoding genes.

4.3 Materials and Methods

Germplasm

The lines evaluated included 597 diverse inbred lines from the Wisconsin Diversity (WiDiv) association panel (Mazaheri et al., 2019), along with four *bm1* RNAi lines and their corresponding wild type (WT) control (Supplemental File 4.1). The WiDiv lines have previously been categorized into subpopulations through admixture analysis (Mazaheri et al., 2019), and all references made to a particular subpopulation in this study used the classifications given there. The subpopulations included in Mazaheri et al. (2019) represent the three major heterotic groups of North American field corn, the Stiff Stalk (SS), Non-Stiff Stalk (NSS), and Iodent (IDT), as well as sweet corn, popcorn, and non-temperate tropical lines. The SS and NSS groups are further sub-divided into additional subpopulations based on heterotic sub-groups. The SS sub-groups include SS-BSSSC0, SS-B14, SS-B73, and SS-B37, and the NSS groups include NSS-Mo17 and NSS-Oh43. Importantly, among the SS sub-groups, SS-BSSSC0 lines (Stucker & Hallauer, 1992) represent relatively unselected maize germplasm from the early 20th century, while the remaining SS sub-groups represent inbreds that maize breeders derived through selection from the SS-BSSSC0 lines. Many of these selected inbreds were previously commercialized under the Plant Variety Protection Act and are classified as ex-PVPs. Similarly, many lines belonging to the NSS-Mo17 and NSS-Oh43 sub-groups are ex-PVPs (White et al., 2020).

Phenolic metabolite data

The experimental design and metabolite data generation are described in Rodriguez et al. (2022), and can also be found in the Materials and Methods section of Chapter 3 of this thesis. Briefly, three replicates (reps) of each inbred line were grown under controlled conditions (21℃, 16 h light and 8 h darkness). Each rep consisted of an independent randomization of all WiDiv lines, and reps were grown at three separate dates. From each plant, the basal 6 cm of stem tissue was harvested from 21-day old seedlings and used for targeted liquid chromatography-mass spectroscopy (LC-MS) metabolic profiling. Methanol extracts were prepared for the three reps, and they were run sequentially in time using a Waters ACQUITY TQD Tandem Quadrupole UPLC/MS/MS (Waters Corporation, Milford, USA).

We downloaded the processed metabolite data from Rodriguez et al. (2022) from the CyVerse Data Commons at https://doi.org/10.25739/e1bq-kh07. It consisted of a matrix of 28 phenolic compounds profiled across the 3 reps of the 602 inbred lines (~1800 x 28 data points). Notably, the units of the LC-MS data are relative values that are normalized to the dry weight of sample used for extracts and expressed as arbitrary units of area (AUA). Although we did not have absolute quantifications, our assumption was that peak area and compound accumulation were proportional to one another, and therefore, we used these terms interchangeably in this study. Our investigation focused on a subset of four phenolic compounds that included coniferyl aldehyde, ferulic acid, sinapic acid, and tricin, which represented key secondary metabolites relevant to lignin biosynthesis.

Metabolite data analysis

We used the following linear mixed model: $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ to characterize the standing genetic variability for each of the four lignin-related phenolic compounds, where Y_{ij}

represented the response variable of the i^{th} genotype in the j^{th} rep, μ was the grand mean, α_i was the effect of the i^{th} genotype, β_j the effect of the j^{th} rep, and ε_{ij} was the residual error. We treated all factors as random effects, with the exception of μ, which we treated as a fixed. To test for normality (both raw data and model residuals), we conducted a Shapiro-Wilk test (W) (Shapiro & Wilk, 1965)**.** Our experimental design allowed us to estimate several sources of variation, including genotypic variance (GenoVar), replication variance (RepVar), and error variance (ErrVar). The repeatability of each compound was calculated as the ratio of GenoVar to the total variation observed. To assess the significance of each term in the model, we used a Chi-square distribution with the REML likelihood ratio (LR) as the test statistic with a type 1 (alpha) error threshold of 5%. As an additional test for cases where there was a significant rep effect in the mixed model, we compared all pairwise differences between means of reps using an unpaired T-test with an alpha of 0.05 for each comparison and calculated all pairwise Spearman rank correlations between reps. If the rep effect was significant for any given pair of reps, but the Spearman rank correlation was also significant, we proceeded with calculating best linear unbiased predictors (BLUPs) from the linear mixed model. To determine the extent to which the four phenolic compounds differentiated lines by subpopulation, we performed a principal components analysis (PCA) on the phenolic compound BLUPs, and visually inspected the groupings by plotting the first two principal components (PCs).

We used several R packages to analyze the metabolite data, including *lme4* (Bates et al., 2015) for model fitting, lmerTest (Kuznetsova et al., 2017) for mixed model significance testing, and ggpubr (Kassambara, 2020) for pairwise mean comparisons and plotting.

Gene expression and genotypic data

We obtained single nucleotide polymorphism (SNP) marker data for the WiDiv lines from Mazaheri et al. (2019) to perform the GWAS analysis and gene expression data from (P. Zhou & Springer, 2020) for conducting the TWAS analysis. The SNP and gene expression data were both generated from the same RNA-Seq reads produced from total RNA extracted from whole seedlings at the V1 developmental stage as described in (Mazaheri et al., 2019). We filtered the SNP data to retain markers with a minor allele frequency (MAF) greater than 0.05, which resulted in a filtered matrix consisting of 428,287 SNPs by 597 WiDiv lines. To filter and normalize the expression data, we used the default GTEx data normalization pipeline with default parameters (Broad Institute, 2022) which retained only protein-coding genes with more than 0.1 transcripts per million (TPM) and equal or more than 6 read counts in more than 20% of the WiDiv lines, and then applied a trimmed mean of M (TMM) normalization. The resulting filtered gene expression data consisted of 26,095 genes by 597 WiDiv lines. The gene expression and SNP data were both based on coordinates from the B73v4 reference genome (Jiao et al., 2017).

Gene annotations and ontologies

To aid with our interpretation of genes identified in this study, we used existing gene annotation and gene ontology (GO) databases. The specific files used were the general feature format file (GFF) for the B73v4 reference downloaded from http://ftp.ensemblgenomes.org/pub/plants/release-48/gff3/zea_mays/ and a GO term to gene ID mapping file (Zea_mays.MSU.GO.terms.gene.txt) published in Hoopes et al. (2018).

Chromosomal binning and covariates

With the MAF filtered SNP data we generated chromosomal bins by splitting each chromosome into 100 quantile-based bins, where the quantiles were based on the physical position of SNPs. Also with the SNP data, we conducted PCA on a set of 121,584 linkage disequilibrium (LD) pruned SNPs and extracted the top five PCs to use in the GWAS analysis as covariates. For LD pruning we used the snp_autoSVD function from the bigsnpr R package (Prive et al., 2018), with the window size argument set to 200 SNPs and the argument pertaining to the maximum squared correlation threshold between SNP pairs set to 0.5. Additionally, we used the first two SNP-derived PCs to compare subpopulation differentiation to PCs generated from PCA on the metabolite data. From the expression data, we extracted the top five probabilistic estimation of expression (PEER) factors (Stegle et al., 2012) using all default parameters from the GTEx data normalization pipeline (Broad Institute, 2022) to use as covariates in the TWAS analysis.

GWAS and TWAS analyses

We performed two separate association analyses to dissect the genome-wide genetic (GWAS) and transcriptional (TWAS) architectures of each phenolic compound using the phenolic compound BLUPs as the dependent variable in each analysis.

For the GWAS analysis, we implemented a modified version of the multi-covariate and aggregation approach depicted in Supplemental Figure 2.2 and described in the Materials and Methods section of Chapter 2 of this thesis. Briefly, we fit a series of seven GWAS models for each phenolic compound, with each model including a progressively increased level of correction for kinship (K) and population stratification (Q), then aggregated the association results based on the persistence of a significant ($p \leq 1/M$, where M was the number of markers tested) association in at least two of the GWAS models fit (Supplemental Figure 2.2). We used the fixed and random model circulating probability unification (FarmCPU) GWAS method which calculated K as part of the iterative procedure and accepted Q as SNP-derived PCs (Liu et al., 2016). The only modification we made to the approach described in Supplemental Figure 2.2 was with respect to the window size that we used (500 kb) for identifying putative candidate genes. Our reasoning for selecting a relatively liberal window size was that LD patterns vary throughout the genome, and since one of our goals was to determine concordance between genes identified by GWAS and TWAS, a smaller window size would have potentially missed shared associations.

To perform TWAS, we also relied on a modified version of the multi-covariate persistence and aggregation approach (Supplemental Figure 2.2). Rather than using the FarmCPU method, however, we used the unified mixed-model method (MLM) and provided Q as PEER factors along with K calculated on the full set of 428,287 SNP markers (Yu et al., 2006). The reason for using the MLM model for TWAS over FarmCPU was that FarmCPU relies on LD in the genome during the iterative process, and while neighboring genes may at times have similar expression patterns, they do not necessarily capture LD in the same way as SNP markers (Williams & Bowles, 2004). In contrast to GWAS where we leveraged LD and scanned up and downstream in the genome from the associated SNP, with TWAS we tested for the association with the expression of each gene independently, without up or downstream genomic searches. When aggregating results across TWAS models fit, we used 1/G as the significance threshold where G represented the number of genes tested.

We designated putative candidate genes as high confidence candidates if they were identified in both the GWAS and TWAS analyses. Furthermore, we emphasized associations where there appeared to be a three-way association (e.g., metabolite-SNP, metabolite-gene expression, and gene expression-SNP).

Generating *bm1* RNAi lines

We developed an RNAi construct to target the *bm1* gene in maize by cloning a synthesized *bm1* sequence (GRMZM5G844562; AGPv3 B73 reference) into the pDONOR207 entry vector and subsequently sub-cloning into the pANIC8D expression construct (Mann et al., 2012). Using our construct (Supplemental Figure 4.1), the staff at the Wisconsin Crop Innovation Center transformed immature maize embryos from the near-isogenic line FBLL-MAB (WT) (Lowe et al., 2006) via *Agrobacterium*-mediated transformation, to produce eight independent *bm1* RNAi events (T_0 lines). At the T_0 generation, we backcrossed all lines to the WT to produce T_1 seed in the greenhouse, followed by two successive generations of self-crossing (field then greenhouse) to produce T_2 and T_3 seeds (Supplemental Figure 4.2). The construct used for transformation included a red fluorescent protein (RFP) selectable marker, which we used to confirm the allelic state (hemizygous or homozygous) of seeds corresponding to each RNAi event. The confirmation was based on the segregation ratios of seeds from individual ears of plants at the T_3 generation, where ears that were not segregating (all RFP positive) were homozygous, and those with a 1:2:1 ratio were assumed to be hemizygous (Supplemental Figure 4.2). Thus, with the known allelic states of plants at the T_3 generation, we returned to the corresponding seed from the T_2 generation to use in this experiment (Supplemental Figure 4.2). Of the eight RNAi events, we chose two to evaluate based on the visible presence (although faint) of the expected *bm1* phenotype and a strong RFP expression in seeds at each generation.

Differential phenolic metabolite accumulation analysis of *bm1* RNAi lines

For each of the two *bm1* RNAi events retained, we evaluated three reps (separate plants) of the hemizygous (E1-He, E2-He) and homozygous (E1-Ho, E2-Ho) allelic states along with the WT. These plants were grown together with the WiDiv experiment described in the phenolic metabolite data section above, and the same tissue (basal 6cm stem) was harvested at the same developmental stage (21-day old seedling) and processed at the same in the same way as WiDiv lines. To differentiate between RNAi allelic states, hereafter we referred to each allelic state as an independent experimental unit (RNAi line). We tested for differential accumulation of each phenolic compound between the WT and each RNAi line by using the biological rep data and conducting an unpaired T-test with an alpha of 0.05 in each test.

To provide context for compounds that we identified as significantly differentially accumulated in at least one RNAi line, we compared the metabolic accumulation patterns in RNAi lines to those observed in the WiDiv as a whole. Additionally, we identified WiDiv lines that either most closely resembled (most-like) or least closely resembled (least-like) each *bm1* RNAi line and the WT. The metric used to identify the most-like and least-like WiDiv lines was based on the minimum and maximum Euclidean distances between points in the PC bi-plot (first two PCs) calculated from BLUPs of the metabolite data.

RNA-Seq and differential gene expression analysis of *bm1* RNAi lines

From the same plants and stem tissue used to generate the LC-MS phenolic metabolite data for the WT and *bm1* RNAi lines, we extracted total RNA using a *Quick*-RNA microprep kit (R1050 ZYMO Irvine, CA) and then treated it with DNase (AM2238 Thermo Fisher Scientific, Waltham, MA). We delivered the purified total RNA to the University of Wisconsin-Madison Gene Expression Center where they prepared the stranded total RNA libraries using the Illumina TruSeq Stranded Total RNA Plant kit (20020596 Illumina, San Diego, CA), and sequenced the libraries on the NovaSeq 6000 platform to generate 151 bp paired-end reads. The number of reads per sample along with the percent GC and percent of unique reads are included in Supplemental Table 4.1.

To analyze the RNA-Seq reads, we used the RASflow Snakemake RNA-seq analysis pipeline with all default parameters (Zhang & Jonassen, 2020). The workflow managed read trimming, alignment, quantification, and DEA. The tools and reference genome data file used by the workflow included TrimGalore for read trimming (Martin, 2011), Salmon (Patro et al., 2017) for read alignment and quantification using the B73 v4 transcriptome downloaded from http://ftp.ensemblgenomes.org/pub/plants/release-48/fasta/zea_mays/cdna/, and lastly, DESeq2 (Love et al., 2014) for conducting the DEA. A summary of the reads processed and mapped per sample is included in Supplemental Table 4.1. Genes with low expression (less than six read counts) across all samples were filtered out in the DEA, and we used the Bonferroni correction threshold ($-\log_{10}$ (p-value) = 5.69) to identify significant differentially expressed (DE) genes. Prior to aggregating the DEA results, we checked if the expression of *bm1* had in fact been significantly reduced in the RNAi lines as we would have expected. We did this using an unpaired T-test with an alpha of 0.05 and excluded the DE results from any of the RNAi lines without a significant reduction in the expression of *bm1* from any downstream analyses.

To filter genes from the DEA, we conducted a gene co-expression analysis for the gene (Zm00001d015618, abbreviated ZmCAD3) which when mutated produces a *bm1*, and the six additional gene members from the same gene family (ZmCAD). The goal of this analysis was to produce a filtered gene set with potentially similar biological roles as ZmCAD genes, by relying on the principle of guilt-by-association (Oliver, 2000). The idea was that if we identified a gene as DE, and it was strongly co-expressed with at least one of the ZmCAD genes, our confidence in the DE result would be higher than basing it strictly on the significance of the test statistic returned by DESeq2. However, for the co-expression analysis to be meaningful, we needed a dataset with larger sample size and more variability than the expression data from the RNAi lines alone. Therefore, we used the same WiDiv gene expression data that we used for TWAS. To assess the strength of the co-expression between each of the ZmCAD genes versus all genes tested, we used the mutual ranking of Pearson correlation coefficients (PCC) between gene pairs (Obayashi & Kinoshita, 2009) and designated the top 5% of values per ZmCAD gene as significant. All analyses pertaining to co-expression were done using R and the tidyverse R package (R Core Team, 2020; Wickham et al., 2019).

4.4 Results

Natural variation for phenolic metabolite accumulation

There was significant (Pr>LR <0.05) genetic variation for each of the compounds we analyzed, with moderate to high estimates of repeatability (Table 4.1). The compound with the highest repeatability was tricin (0.77), whereas the lowest repeatability (0.64) was for coniferyl aldehyde (Table 4.1). Repeatability estimates for sinapic and ferulic acid were 0.68 and 0.74, respectively (Table 4.1). The peak area distributions by replication deviated from normality for all compounds (Pr<W<0.05), and there was a notable right skew in the distributions of coniferyl aldehyde, ferulic acid, and sinapic acid (Figure 4.1). However, for downstream analyses such as GWAS and TWAS, the assumptions of normality are not based on raw data distributions, but rather on residuals, which were more normally distributed than the raw distributions (Figure 4.2).

We intended to combine the data across biological replicates for downstream analyses, however, there was a significant (Pr>LR <0.05) replication effect for coniferyl aldehyde and ferulic acid (Table 4.1, Figure 4.1). When evaluating the three replicates for these two compounds, we did not identify any cases where all between replication comparisons were significant (Figure 4.1). Furthermore, the Spearman rank correlations between replicates were all predominantly linear and significant (Figure 4.1). Therefore, we proceeded with calculating

BLUPs for all compounds, which provided us with a single value per compound to represent each inbred line (Figure 4.1).

PCA on the phenolic compound BLUPs explained 70.2% of the observed metabolite variance with the first two PCs, while the first two PCs calculated on the LD pruned SNP data explained 8.98% of the marker variation (Figure 4.3). These metabolites and SNP-based dimension reduction analyses were meant to determine to what extent one subpopulation of related individuals might be enriched or fixed for a certain feature (SNPs or metabolites). The PC bi-plot of the SNP data very clearly differentiated subpopulations from one another, particularly the NSS-Mo17, SS-B73, and IDT. In contrast, the PC bi-plot corresponding to the metabolite data did not (Figure 4.3). This would suggest that the metabolite variation among subpopulations was on par, or less than the variation observed within subpopulations.

Genomic regions and candidate genes associated with phenolic metabolite accumulation

In the GWAS analysis, we identified a total of 85 significant (-log10(p) > 5.63) SNP-metabolite associations across the four focus compounds (Figure 4.4, Supplementary Figure 4.3). The number of significant SNPs was approximately equally distributed among the compounds, with 26 corresponding to ferulic acid, 22 to tricin, 19 to sinapic acid, and 18 to coniferyl aldehyde (Figure 4.4). No single SNP was significantly associated with more than one compound, however, there were overlaps in terms of the chromosomal bins (Supplementary Figure 4.4) to which the SNPs physically mapped to. For example, SNPs associated with ferulic acid and sinapic acid were within the same 3.9 Mb bin on chromosome 1, and SNPs associated with coniferyl aldehyde and tricin were within the same 2.2 Mb and 1.2 Mb bins on chromosomes 3 and 4, respectively (Figure 4.4). Similarly, a 3.8 Mb region on chromosome 5 harbored significant associations for all compounds, including the most significant association for tricin previously mentioned (Figure 4.4). The SNP-to-gene assignment yielded a total of 2,357 genes, many of which corresponded to GO molecular function and biological processes terms related to molecular binding, transmembrane transport, oxidation and reduction processes, and protein phosphorylation (Supplementary Figure 4.5).

The advantage of the TWAS analysis in terms of identifying candidate genes was that we were directly testing for an association between each phenolic compound and the expression of the genes themselves. With a relatively loose threshold of -log10(p) = 2.00, we identified a total of 443 protein-coding genes whose expression was associated with the relative abundance of at least one phenolic compound (Figure 4.4). The majority of genes identified were associated with ferulic acid (154 genes) and coniferyl aldehyde (149 genes), whereas fewer were identified for sinapic acid (84 genes) and tricin (70 genes). There were 14 genes that were associated with more than one compound, the majority of which were shared between ferulic acid and sinapic acid (Figure 4.4). The GO molecular function and biological process terms corresponding to the 443 TWAS identified genes largely mirrored the GO terms for the GWAS identified genes (Supplementary Figure 4.5). The shared GO annotations provided a general sense of agreement between the two analyses, however, it alone was not a direct gene-to-gene comparison.

For each compound, we compared the putative candidate genes identified by GWAS to those identified by TWAS to generate a list of high confidence associations that were concordant between the two analyses. We identified a total of 20 genes that met this criterion, of which, nine corresponded to ferulic acid, seven to coniferyl aldehyde, three to sinapic acid, and only one for tricin (Table 4.2, Figure 4.4). Two of the high confidence candidate genes for coniferyl aldehyde (Zm00001d030138 and Zm00001d053438) and one for sinapic acid (Zm00001d028008) harbored the SNPs associated with the compound within their respective gene start and end sites, suggesting those SNPs might be tagging the causal gene, and possibly be associated with the expression of the gene as well. The connection between a local SNP (within the gene start and end site), gene expression, and metabolite accumulation was something we observed for Zm00001d030138 and Zm00001d028008, which were associated with coniferyl aldehyde and sinapic acid, respectively (Figure 4.5).

For the coniferyl aldehyde association with Zm00001d030138, inbred lines with the highest accumulation of coniferyl aldehyde tended to have low expression of Zm00001d030138, and those with low gene expression generally carried the reference (B73 v4) allele (Figure 4.5). The gene annotation for Zm00001d030138 was listed as a probable NADH:ubiquinone dehydrogenase (Table 4.2), and its ortholog in Arabidopsis (At3G03100) has been characterized as encoding a mitochondrial complex I subunit (Heazlewood et al., 2003), but to our knowledge, no characterization of this particular gene has been done in maize.

With respect to the association of Zm00001d028008 with sinapic acid, individuals with a high expression of Zm00001d028008 predominantly carried the reference (B74 v4) allele which in turn was associated with the highest levels of sinapic acid accumulation (Figure 4.5). Similar to Zm00001d030138, the gene annotation for Zm00001d028008 (probable mitochondrial adenine nucleotide transporter) also suggested a role in energy metabolism (Table 4.2). Though no formal characterization of the gene has been done in maize, in rice, the protein encoded by the ortholog of Zm00001d030138 (Os03g09110) was found to have a reduced accumulation in germinating seedlings as a response to nitric oxide stress (Mao et al., 2018).

Perturbed expression of *bm1* by RNAi

The *bm1* gene did not appear in any of the GWAS or TWAS analyses. However, based on its known function in reducing coniferyl aldehyde, sinapyl aldehyde, and *p*-coumaryl aldehyde to their respective monolignol alcohols (Mansell et al., 1976; Morrison et al., 1994), we used it as a means to investigate the metabolic and transcriptomic consequences of artificially disrupting its expression by RNAi. Of the *bm1* RNAi lines, E2-He and E2-Ho had inconsistent and non-significant (p>0.05) reductions in the expression of *bm1* compared to the WT and therefore we excluded them from all downstream analyses (Supplementary Figure 4.6). In contrast, *bm1* expression compared to the WT was significantly (p<0.05) reduced in E1-He and E1-Ho by 53.8% and 61.3%, respectively (Supplementary Figure 4.6).

Differentially accumulated metabolites in *bm1* RNAi lines

Compared to the WT, we observed a significant (p<0.05) increase in the accumulation of all compounds in at least one of the *bm1* RNAi lines retained (Table 4.3). The compound with the largest increase was coniferyl aldehyde, with a nearly 9-fold increase over the WT for E1-Ho lines, and a 3-fold increase over the WT for E1-Ho lines (Table 4.3). The 9-fold increase for E1-Ho lines was highly significant (p<0.01), whereas the 3-fold increase corresponding to E1-He lines did not surpass the predefined 0.05 alpha threshold (p = 0.08) (Table 4.3). The increases in ferulic acid and sinapic acid over the WT were significant (p<0.05) for both E1-He and E1-Ho lines, and the lines accumulated approximately two times as much ferulic acid as the WT and three times as much sinapic acid (Table 4.3). E1-Ho lines alone had a significant (p<0.05) increase in tricin accumulation, about 75% higher than the WT (Table 4.3). These results were consistent with what has been described for a similar set of compounds in mature (silage stage) maize stalk internodes from mutant *bm1* maize plants (Liu et al., 2021), suggesting that phenolic compound accumulation (in stems) at the seedling stage may persist through development, and thus might be useful for predicting lignin composition or related traits such as saccharification in mature plants.

To contextualize the metabolic profiles observed for the *bm1* RNAi lines, we compared them to the natural variation observed among lines from the WiDiv. Based purely on the distance between points on the first and second PCs calculated from the metabolite data BLUPs, we identified the top five WiDiv lines (~1%) that most closely resembled each of the RNAi lines and those that least resembled the RNAi lines (Figure 4.6, Table 4.4). The WiDiv lines most-like E1-He clustered more tightly than those most-like E1-Ho, which emphasized the distinct profile that E1-Ho possessed (Figure 4.6). The uniqueness of E1-Ho was exemplified particularly by the combination of high coniferyl aldehyde and tricin that was well beyond the range in natural variation observed in the WiDiv (Figure 4.6). Among the top five WiDiv lines most-like the WT, three belonged to the SS-BSSSC0 group, and among the top five WiDiv lines most-like either E1-Ho and E1-He 3 lines belonged to the SS-B14 group (Table 4.4).

Differentially expressed genes in *bm1* RNAi lines

We identified a large number of significantly ($-\log_{10}$ (p-value) ≤ 5.69) DE genes in both E1-Ho (1,167 genes) and E1-He (823 genes), most of which were downregulated (76.9% of E1-Ho DE genes and 78.4% of E1-He DE genes) (Supplemental File 4.2). To filter the DE genes, we performed a genome-wide co-expression analysis for each of the ZmCAD gene family members and additionally retained any of the high confidence GWAS/TWAS genes (Table 4.2).

The co-expression patterns varied between ZmCAD gene family members, with the majority having predominantly positive PCCs with their top 5% of co-expressed genes, and only two (ZmCAD2, ZmCAD5) having roughly equal proportions of positive and negative PCCs (Figure 4.7). The co-expression patterns for *bm1* were notable due to the magnitude of the PCC values compared to the other gene family members (Figure 4.7). Among the most stringently filtered (top 0.05%) co-expression results, we identified several known phenylpropanoid biosynthesis-related genes including Zm00001d049541 (*bm*3) and Zm00001d015459 (*bm*5), which were both co-expressed with *bm*1, and similarly, confer a brown-midrib phenotype when their expression is disrupted in mutant plants (Vignols, 1995; Xiong et al., 2020). These results suggested that our approach of using co-expression as a filter for DE genes did indeed have merit.

After filtering the DE genes based on co-expression, we retained 347 significantly ($-\log_{10}$ (p-value) ≤ 5.69) DE genes for E1-Ho and 242 genes for E1-He (Figure 4.8). Among the filtered DE genes, we identified an HCALDH gene (Zm00001d045706) that was DE in both E1-Ho and E1-He (Table 4.5, Figure 4.8). Our expectation based on previous work was that HCALDH would have been upregulated to convert excess coniferyl aldehyde accumulation due to the *bm1* disruption to ferulic acid (Mittasch et al., 2013; Nair et al., 2004). However, we instead observed an approximately equal downregulation (-0.72 log2-FC) in both E1-He and E1-Ho (Table 4.5). The alternative path to ferulic acid, suggested by (Liu et al., 2021), was through a potential indirect path by a protein interaction between CAD and the upstream CCR. Interestingly, we observed that all CCR genes that were DE were downregulated (Table 4.5). If gene expression of CAD and CCR encoding genes correlates with protein abundance, our results would be in line with the CCR and CAD interaction hypothesis made by Liu et al. (2021).

Comparing the DE genes to the high confidence GWAS and TWAS associations, we identified overlaps with three genes (Table 4.2, Table 4.5). These included genes associated with coniferyl aldehyde (Zm00001d051675), ferulic acid (Zm00001d003016), and sinapic acid (Zm00001d045454), and they were all upregulated in the *bm1* RNAi lines. The gene associated with sinapic acid accumulation (Zm00001d045454) was the only gene that was DE in both the E1-Ho and E1-He lines (Table 4.5, Figure 4.8).

4.5 Discussion

Efforts to understand the connection between genotype and phenotype for complex traits have historically relied on DNA markers to associate with end use and observable phenotypes. Technological advancements in data generation at multiple molecular levels including the transcriptome, metabolome, and proteome now present unique opportunities to dissect the path from genotype to phenotype (Y. Yang et al., 2021). In maize, previous studies have demonstrated the utility of integrating multiple omics data types to predict complex traits at both the inbred and hybrid levels (Azodi et al., 2020; Guo et al., 2016; Schrag et al., 2018; Westhues et al., 2017). While extremely useful in an applied sense, these prediction-based studies focused primarily on establishing statistical frameworks and did not necessarily provide much insight into the underlying biology.

Here, we leveraged the natural genetic variability of the WiDiv to dissect the molecular trait architecture of four key phenolic compounds that serve as important intermediates and components of the lignin biopolymer (del Río et al., 2020). Our evaluation of lignin-related metabolites in stems from 21-day old seedlings may have seemed unusual given that typically lignin-focused studies evaluate plant tissues that have undergone secondary cell wall formation and are heavily lignified, such as the maize stalk (Hansey et al., 2010; Jung et al., 1998). However, given the wide range in developmental rates among many WiDiv lines (Hirsch et al., 2014), evaluating earlier in development allowed for greater control over this potential confounder. Furthermore, we speculate that some determinants of phenolic compound accumulation may be conserved through development, but would potentially be muddled by inabilities to control for increased experimental error at larger scales such as in field conditions. Lastly, the largest number of genes are expressed at the seedling stages (Sekhon et al., 2011).

Therefore, by evaluating our materials at nearly the same developmental stage as was used to generate the WiDiv expression data, we maximized our potential to associate phenolic profiles with gene expression profiles that were representative.

Unlike SNPs, which do not have spatial or temporal variation, gene expression and metabolite accumulation are inherently dynamic and can vary between plant tissues, through development, and are responsive to both external and internal cues (Hoopes et al., 2019; Othibeng et al., 2021; S. Zhou et al., 2019). Among the phenolic compounds analyzed, coniferyl aldehyde had the lowest repeatability (0.60) estimate (Table 4.1). This would suggest that either the targeted LC-MS method for detecting coniferyl aldehyde had more instrumental error associated with it, or that coniferyl aldehyde accumulation in seedlings is sensitive to even small environmental condition differences since all plants were evaluated under the same controlled conditions.

In our TWAS analysis, we used early development (V1) non-tissue specific (whole seedling) gene expression data from the WiDiv to associate with the 21-day old (~V3) phenolic compound accumulation data. When associating traits, or metabolites in our case, to gene expression data that is not from the same tissue or developmental stage, or the plant being evaluated, there is always a concern that the expression profile used may not fully reflect the expression profile from when the evaluation was performed. A potential follow-up confirmation could be to perform RNA-Seq on the same tissue used for LC-MS, since only a small (50 mg) portion of the stem material was consumed for the metabolic profiling. However, a more strategic and cost-effective approach might entail choosing selected inbred lines for RNA-Seq that either represent metabolic extremes, such as those listed in Table 4.4, or select lines that have contrasting gene expression profiles which drive the associations for the high confidence candidate genes, such as those depicted in Figure 4.5.

Our study was not the first to integrate GWAS and TWAS in maize (Hirsch et al., 2014; Kremling et al., 2019) or to identify candidate genes associated with the accumulation of phenylpropanoid compounds (S. Zhou et al., 2019). What differentiated our study was our specific targeting of phenolic compounds related to lignin biosynthesis and our distinct approach to combining GWAS and TWAS analyses. In the Kremling et al. (2019) study, the authors assigned SNP associations from GWAS to the nearest gene (based on physical distance) and then used a Fisher's combined test (Fisher, 1938) to combine the GWAS and TWAS associations pertaining to the same gene (Kremling et al., 2019). Although LD decays rapidly in maize (Wallace et al., 2014), assigning SNPs to only the nearest gene would ignore variable, and more importantly, longer-range regions of LD in the maize genome (Flint-Garcia et al., 2003), potentially missing shared associations between GWAS and TWAS. In contrast, in our approach, we used a liberal window size of 500 kb to assign metabolite-associated SNPs to all genes within their respective windows, which led us to identify 20 high confidence candidate genes across the four phenolic compounds (Figure 4.4, Table 4.2). A potential extension to our comparative-based GWAS and TWAS concordance method could come from utilizing a recently proposed meta-analysis type aggregation approach to unify results (Yoon et al., 2021).

Our differential metabolite and gene expression analysis of *bm1* RNAi lines complemented our association analyses and allowed us to test specific hypotheses generated from previous work that investigated a similar set of phenolic compounds in mutant *bm1* lines (Liu et al., 2021). Between DEA, GWAS, and TWAS we identified three genes that overlapped in all analyses (Table 4.2, Table 4.5). The gene associated with sinapic acid accumulation

(Zm00001d045454) was the only gene that was DE in both the E1-Ho and E1-He lines (Table 4.5, Figure 4.8), and interestingly the gene annotation of Zm00001d045454 is a zinc finger (C3HC4-type RING finger) family protein, and the CAD enzyme encoded by *bm1* relies on zinc as a cofactor. Together, with the multiple sources of evidence pointing to the same gene association, we can speculate that Zm00001d045454 is likely a crucial factor in modulating the last steps of the monolignol biosynthesis pathway and that there is clearly variation at various molecular levels that can potentially be exploited.

The reasons are unclear as to why we observed an unexpected downregulation in the expression of the HCALDH gene in *bm1* RNAi lines (Table 4.5, Figure 4.8). One possibility is that while there may be sequence homology between the gene sequences of the maize HCALDH and the Arabidopsis and brassica HCALDH genes (Mittasch et al., 2013; Nair et al., 2004), the maize HCALDH genes may have diverged sufficiently and therefore not actually have the same function. An alternative explanation might be that both *bm1* and HCALDH are regulated by a common transcriptional factor, and a reduction in *bm1* initiates a type of regulatory feedback response that reduces the expression of shared target genes. The most likely explanation at least for the increased abundance of ferulic acid is that there is a CAD-CCR interaction occurring in maize similar to poplar (Yan et al., 2019), and the three CCR genes identified might be most actively involved in that interaction (Table 4.5).

Another benefit of including the *bm1* RNAi lines in our study was that we were able to provide context to the magnitude of the effects on phenolic compound accumulation elicited by the perturbation of *bm1* beyond a simple comparison to a WT control. We did this by comparing them to the range in natural variation observed among the WiDiv lines and in doing so we identified several WiDiv lines with *bm1*-like metabolite properties (Figure 4.6, Table 4.4). The presence of several SS-B14 WiDiv lines in the most-like *bm1* group might suggest that maize breeders have indirectly selected for increased phenolic composition within the stiff stalk related germplasm pools. Future research could examine the genome-wide selection signatures associated with phenolic compound accumulation, similar to work that has been done on the WiDiv for tassel morphology (Gage et al., 2018). The elevated phenolic profile observed in the *bm1*-like SS-B14 lines might also be due to our evaluation under high-intensity LEDs. Phenolics have been shown to increase with LED illumination (Lee et al., 2014; Seo et al., 2015), however here all lines were evaluated together, uniformly. Therefore, it is possible that the SS-B14 lines respond to light stress by eliciting a response in phenolic compounds more than in other subpopulations.

We have presented here a large-scale phenolic compound dissection by various association analyses aimed at gene identification and causal inference. Notably, we produced a list of high confidence candidate genes and identified specific inbred lines which can be potentially useful for guiding future breeding efforts for increased phenolic content. In addition, the regulatory genes encoding TF may be useful to guide additional validation studies.

4.6 Tables

Table 4.1: Variance decomposition and BLUP summary statistics of phenolic metabolites from the evaluation V1 whole seedling gene expression data from 597 WiDiv lines in controlled conditions. For each phenolic compound variances of peak areas normalized to dry weight, in arbitrary units of area (AUA) are decomposed into genotypic (GenoVar), replication (RepVar), and error (ErrVar) variances. The repeatability (R) is listed for each compound along with the BLUP mean and standard deviation (SD).

	Variance decomposition				**BLUP summary statistics**
Phenolic compound	**Geno Var**	**RepVar**	**ErrVar**	**R**	**Mean +/- SD**
Coniferyl aldehyde	4.69 ***	0.0629 ***	3.01	0.60	4.29 +/- 1.96
Ferulic acid	6,770 ***	40.40 ***	2,300	0.74	124 +/- 77.9
Sinapic acid	5.27 ***	0.012	2.40	0.68	3.2 +/- 2.13
Tricin	671,000 ***	1,190	190,000	0.77	2820 +/- 782

Table 4.2: High confidence candidate genes identified in both GWAS and TWAS for four phenolic compounds using V1 whole seedling gene expression data from 597 WiDiv lines. Gene IDs with an asterisk denote that the gene was identified as differentially expressed in bm1 RNAi lines as well. Effect sizes are given in terms of number of standard deviations.

Phenolic compound	**Gene ID**	**Chrom**	**Gene TSS (bp)**	**SNP**	**SNP dist. to gene (bp)**	**SNP effect**	**SNP -log10(p)**	**GeneExpr effect**	**GeneExpr -log10(p)**	**Gene annotation**
Coniferyl aldehyde	Zm00001d030138	Chr1	108425304	rs1_108425377	73	0.301	6.27	-0.166	2.43	Probable NADH:ubiquinone dehydrogenase
	Zm00001d034256	Chr1	288168990	rs1_288179712	2,791	0.137	8.20	-0.166	2.37	phosphohexose isomerase1
	Zm00001d034248	Chr1	287871005	rs1_288179712	306,882	0.137	8.20	0.159	2.16	not_annotated
	Zm00001d043037	Chr3	187134578	rs3_187051481	83,097	0.143	6.22	-0.181	2.20	not_annotated
	Zm00001d051675*	Chr4	166355030	rs4_166717644	360,644	0.244	6.37	-0.127	2.33	Transmembrane 9 superfamily member 12
	Zm00001d053438	Chr4	231449438	rs4_231450946	1,508	-0.255	7.12	0.153	2.20	Putative heavy metal transport/detoxification superfamily protein
	Zm00001d026500	Chr10	147239494	rs10_146945635	293,859	0.269	6.84	-0.119	2.03	not_annotated
Ferulic acid	Zm00001d027611	Chr1	8981882	rs1_8597692	384,190	-0.102	6.21	0.103	2.33	Pentatricopeptide repeat-containing protein

	Zm00001d003016*	Chr2	29538173	rs2_29280158	258,015	0.124	10.09	0.147	2.05	phenylalanine ammonia lyase2
	Zm00001d049725	Chr4	41540958	rs4_41735494	189,201	0.144	5.98	-0.142	2.46	GDP-mannose transporter GONST1
	Zm00001d053967	Chr4	244414417	rs4_244519800	101,403	0.111	5.84	-0.115	2.06	Putative bZIP transcription factor superfamily protein
	Zm00001d013212	Chr5	6625547	rs5_6711255	84,179	0.23	8.16	0.118	2.01	Peroxidase 16
	Zm00001d008982	Chr8	28558735	rs8_28575414	14,113	0.141	6.15	0.132	2.94	Reticulon-like protein B4
	Zm00001d010961	Chr8	135100380	rs8_134760788	339,592	0.239	6.88	0.259	3.07	Putative CRINKLY4-like receptor protein kinase family
	Zm00001d012537	Chr8	176319099	rs8_175876486	442,613	0.364	7.48	-0.125	2.47	ADP-ribosylation factor A1F
	Zm00001d047455	Chr9	130773863	rs9_130553775	220,088	0.169	7.88	-0.138	2.15	Alpha-N-acetylglucosaminidase
Sinapic acid	Zm00001d027425	Chr1	4979131	rs1_4833143	145,988	0.214	6.33	0.119	2.10	Agamous-like6 (MADS TF)
	Zm00001d028008	Chr1	20133912	rs1_20134573	661	0.204	6.29	0.147	2.48	Probable mitochondrial adenine nucleotide transporter
	Zm00001d045454*	Chr9	22857763	rs9_23156434	296,279	0.189	7.18	0.121	2.29	Zinc finger (C3HC4-type RING finger) family protein
Tricin	Zm00001d033980	Chr1	279980867	rs1_279923779	57,088	0.101	8.90	-0.117	2.01	ustilago maydis induced12

Table 4.3 Summary of phenolic metabolites profiled in bm1 RNAi lines and the wild type (WT) controls for four phenolic compounds using V1 whole seedling gene expression data from 597 WiDiv lines. Values are the mean peak areas normalized to dry weight, in arbitrary units of area (AUA) of three biological replicates per line +/- the standard deviation (SD). Comparisons are relative to the WT and are provided in fold-change (FC) and percent change.

Phenolic compound	**WT**	**E1-He**				**E1-Ho**			
	Mean +/- SD	**Mean +/- SD**	**P-value**	**FC**	**%Change**	**Mean +/- SD**	**P-value**	**FC**	**%Change**
Coniferyl aldehyde	2.66 +/- 0.541	7.95 +/- 2.81	0.08	2.99	198.9%	23.9 +/- 7.61	0.04	8.98	798.5%
Ferulic acid	71.9 +/- 36	153 +/- 37.7	0.04	2.13	112.8%	171 +/- 44.6	0.04	2.38	137.8%
Sinapic acid	2.3 +/- 1.21	7.19 +/- 0.6	0.009	3.13	212.6%	6.52 +/- 1.25	0.01	2.83	183.5%
Tricin	3010 +/- 361	3590 +/- 485	0.177	1.19	19.3%	5300 +/- 384	0.002	1.76	76.1%

Table 4.4: WiDiv lines with contrasting resemblance to bm1 RNAi plants. The designations are based on using the first two PCs calculated from the metabolite data as a distance matrix and subsetting lines which most or least resemble the RNAi lines and the wild type. Values of each compound listed are BLUPs from peak areas normalized to dry weight, in arbitrary units of area (AUA).

WiDiv Line	**Subpopulation**	**Coniferyl**	**Ferulic**	**Sinapic**	**Tricin**
Most like WT					
CO256	SS-BSSSC0	3.51	82.41	2.41	2193.27
N215	Mixed	4.25	97.28	1.53	1998.94
N523	SS-BSSSC0	5.01	105.83	1.62	1700.64
R227	SS-BSSSC0	3.05	73.11	2.57	2601.58
Va14	Mixed	2.90	86.59	2.08	2720.52
Most like E1-Ho					
B64	SS-B14	8.51	150.59	3.73	4653.60
LH220Ht	SS-B14	8.62	100.81	6.09	4932.94
M37W	Mixed	13.99	76.03	4.61	4011.28
PHK46	Mixed	6.99	101.99	4.68	4984.46
W803G	Tropical	21.88	146.81	2.02	2447.43
Most like E1-He					
B14	SS-B14	4.45	188.99	5.65	3977.28
CG10	Broad origin-public	4.97	115.86	7.75	3743.43
EP1	Sweet corn	4.87	272.47	5.01	4578.02
PHN66	Mixed	7.52	201.07	3.82	3096.60
WIL900	NSS-Mo17	5.98	194.32	5.23	3495.38
Least like E1-Ho and E1-					
52220	Mixed	3.27	38.95	0.46	1478.65
A	Broad origin-public	1.83	93.37	0.39	1930.79
CO125	Broad origin-public	1.68	153.82	5.42	533.66
Ky226	Mixed	3.02	43.59	0.40	1377.30
Mo5	Mixed	1.64	35.87	1.88	981.52
PHV53	NSS-Oh43	2.86	121.63	1.17	860.84

Table 4.5: Summary of the top differentially expressed (DE) genes between the wild type (WT) and bm1 RNAi lines in the hemizygous (E1-He) and homozygous (E1-Ho) allelic states. The top 10 (5 up-regulated and 5 down-regulated) for each RNAi line are listed. Rows shaded in light grey represent DE genes not in the top 10, but rather they are members of either the HCALDH or CCR gene families. Rows in dark grey correspond to high-confidence candidate genes identified in GWAS and TWAS. Individual cells with a dash mean the DE was not significant. The Pearson correlation coefficient (PCC) with bm1 was calculated using V1 whole seedling gene expression data from 597 WiDiv lines.

		E1-He		**E1-Ho**		
Gene ID	**PCC**	**log2-FC**	**-log10(p)**	**log2-FC**	**-log10(p)**	**Gene annotation**
Zm00001d015091	0.527	1.99	29.2	2.19	23.5	Bax inhibitor 1
Zm00001d043419	0.254	3.12	12	1.85	12.4	not_annotated
Zm00001d051420	0.592	2.17	14.8	1.54	15.8	dehydrin3
Zm00001d045454	0.312	1.47	7.3	1.29	9.5	Zinc finger (C3HC4-type RING finger) family protein
Zm00001d033098	0.537	-2.81	57.4	-2.76	53.8	post-illumination chlorophyll fluorescence increase
Zm00001d026606	0.247	-3.61	37.1	-3.53	54	DNAJ heat shock N-terminal domain-containing protein
Zm00001d011183	0.268	-1.89	33.3	-1.8	75.8	thiamine biosynthesis1
Zm00001d045706	0.258	-0.72	6.9	-0.7	11.7	(HCALDH) Restorer of fertility2
Zm00001d020958	0.125	-1.51	2.9	-1.33	7.6	(CCR) Dihydroflavonol-4-reductase
Zm00001d006037	-0.282	-1.51	1.9	-1.75	3	(CCR) NAD(P)-binding Rossmann-fold superfamily protein
Zm00001d034986	-0.092	3.89	21.4	-	-	Cytochrome P450 71A26
Zm00001d025299	0.25	3.04	11.8	-	-	Extensin
Zm00001d023420	-0.434	-	-	2.88	15.4	Protein TERMINAL FLOWER 1
Zm00001d032253	0.218	-	-	2.51	16.1	Type IV inositol polyphosphate 5-phosphatase 9
Zm00001d051675	-0.079	-	-	0.65	11	Transmembrane 9 superfamily member 12
Zm00001d003016	0.701	-	-	0.58	6.1	phenylalanine ammonia lyase2
Zm00001d001960	0.623	-	-	-1.7	44.3	Naringenin2-oxoglutarate 3-dioxygenase
Zm00001d040202	0.334	-	-	-1.94	46.1	SNF7 family protein
Zm00001d012981	0.043	-	-	-1.98	5.3	(CCR) NAD(P)-binding Rossmann-fold superfamily protein
Zm00001d027729	0.818	-3.03	26	-	-	Photosynthetic NDH subunit of subcomplex B 2 chloroplastic
Zm00001d044017	-0.568	-3.59	26.7	-	-	Probable plastid-lipid-associated protein 14 chloroplastic

4.7 Figures

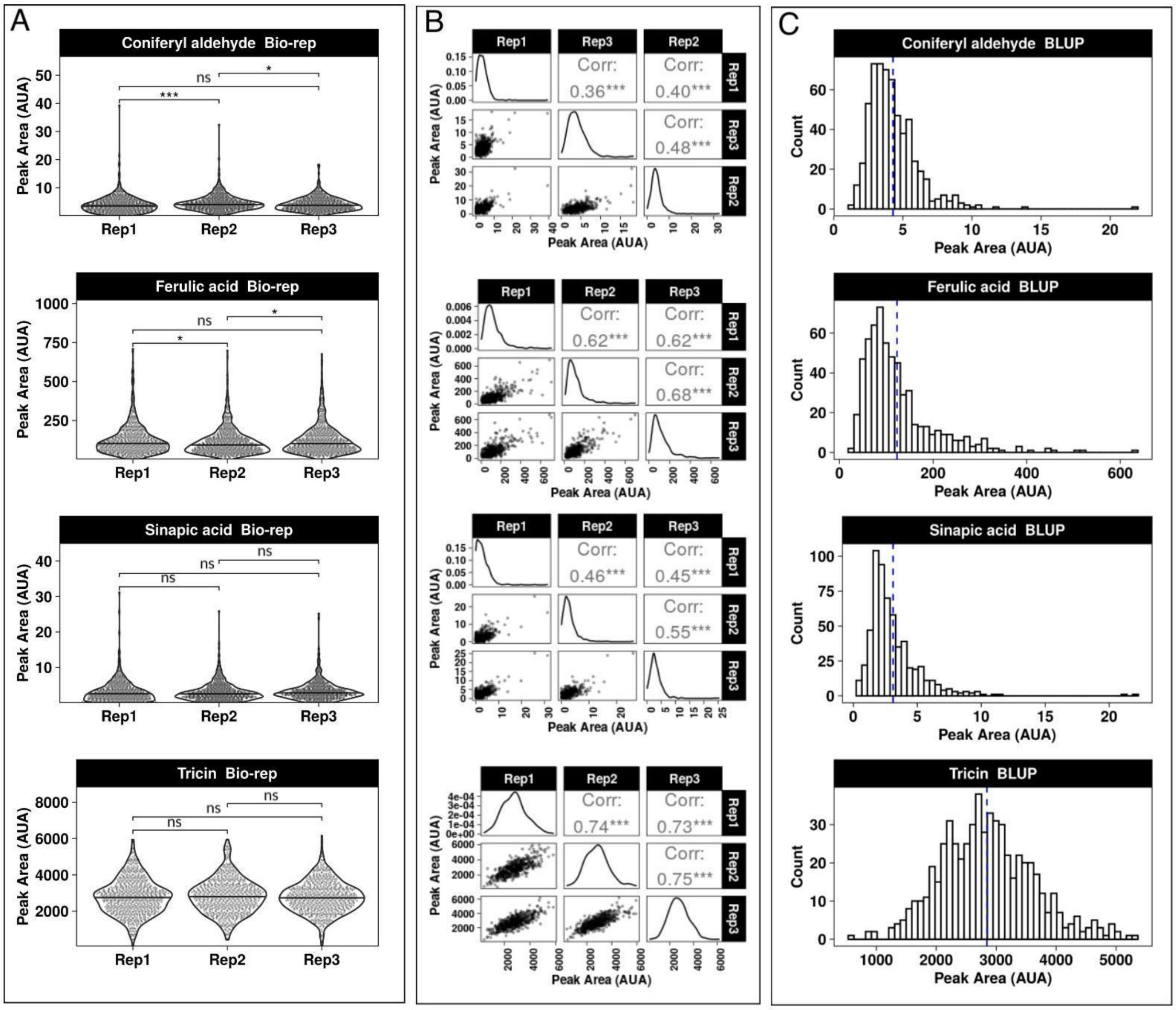


Figure 4.1: Metabolite data analysis across three biological replicates (rep) for 597 WiDiv lines plus four bm1 RNAi lines and the WT (N = 602). Peak areas are relative weight normalzied values and expressed as arbitrary units of area (AUA). (A) Violin distribution plots separated by rep. (B) Between rep Spearman rank correlation (upper diagonal) and bi-plots (lower diagonal). (C) Histograms of BLUPs. Significance codes are from unpaired two-sided T-tests comparing differences between reps, where Pr(>T) <0.01 '**' ; <0.05 '*' ; >0.05 ` `.

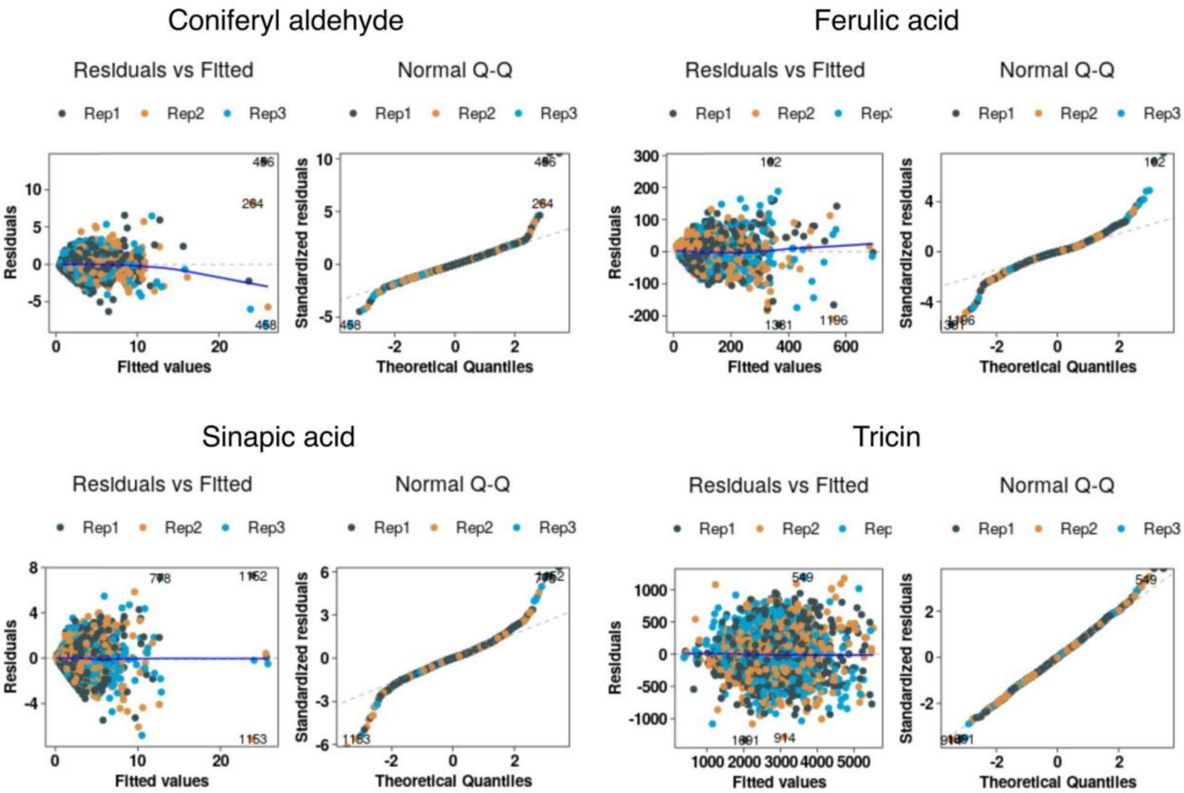


Figure 4.2: Diagnostics plots from the linear mixed model fitted to each of the four highlighted phenolic compounds.

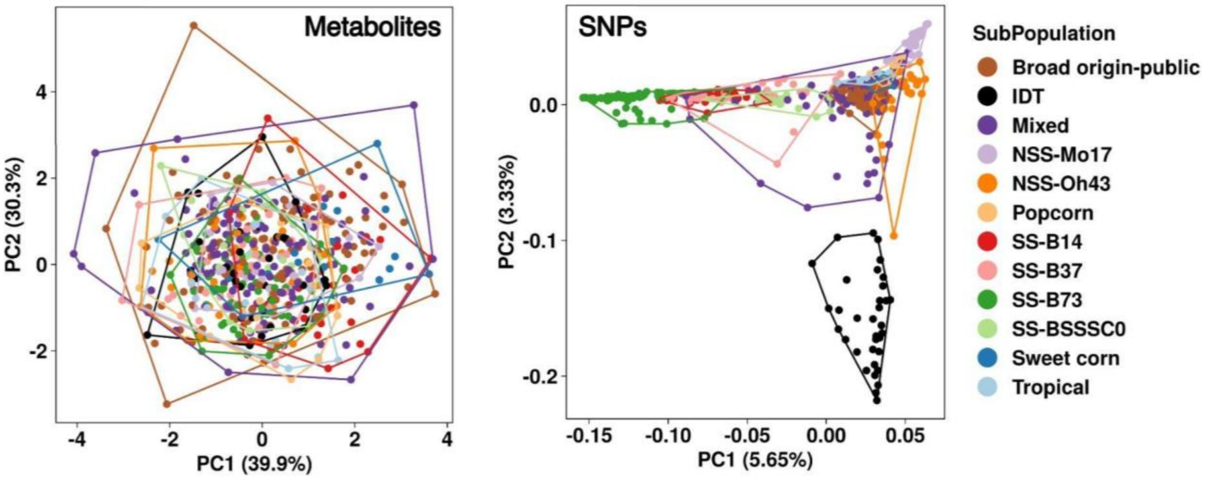


Figure 4.3: The first two PCs calculated from either phenolic compound BLUPs (left panel), or SNPs (right panel) corresponding to 597 WiDiv lines evaluated as V1 whole seedling in controlled conditions. Points are colored by subpopulation, and the surrounding polygon for each subpopulation defines the bounds for members of each respective group.

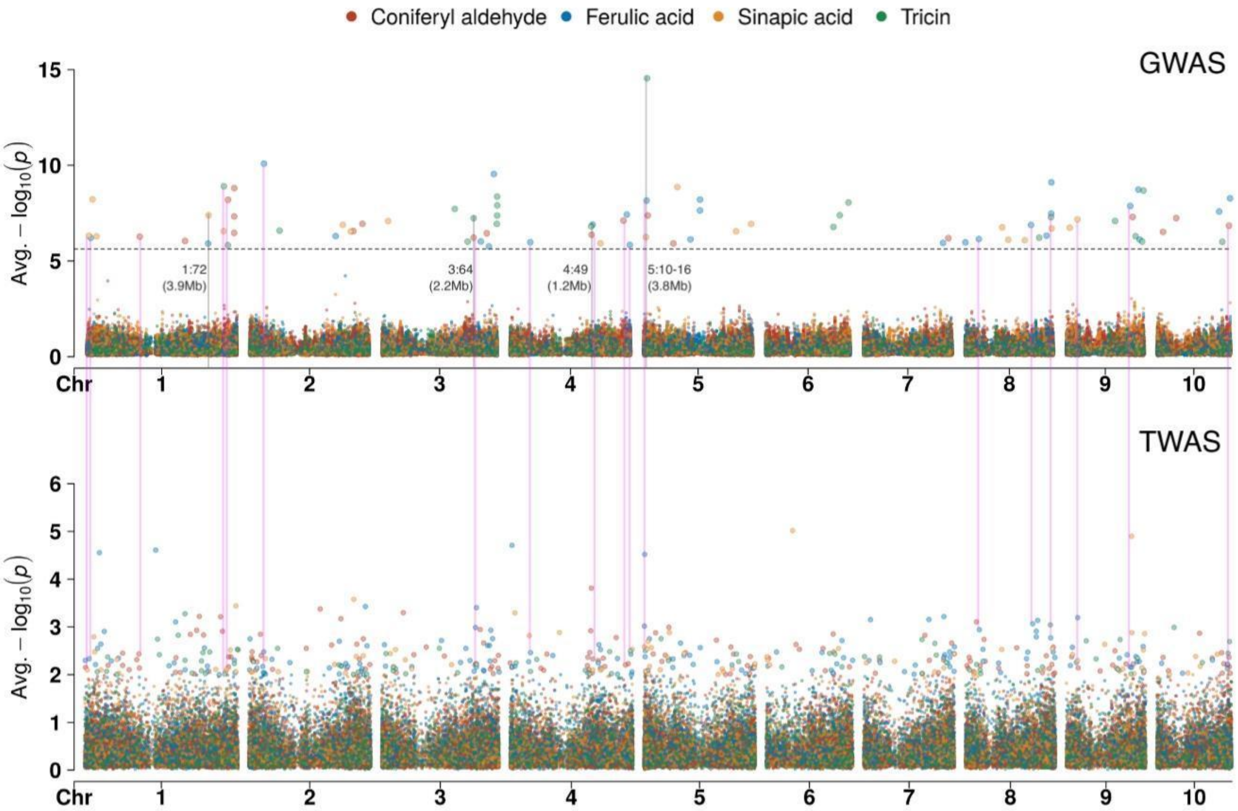


Figure 4.4: Manhattan plots for GWAS and TWAS for metabolite accumulation using V1 whole seedling gene expression data from 597 WiDiv lines. The independent variables tested for an association with each phenolic compound in GWAS were 428,287 SNPs, and for TWAS, the expression of 26,095 protein-coding genes. In the GWAS plot, the annotated regions follow the notation of chromosome:bin and the size of the region is enclosed in parenthesis. The dashed horizontal line in the GWAS plot corresponds to the suggestive threshold (-log10(p) = 5.63), and the pink lines intersecting across plots correspond to shared associations.

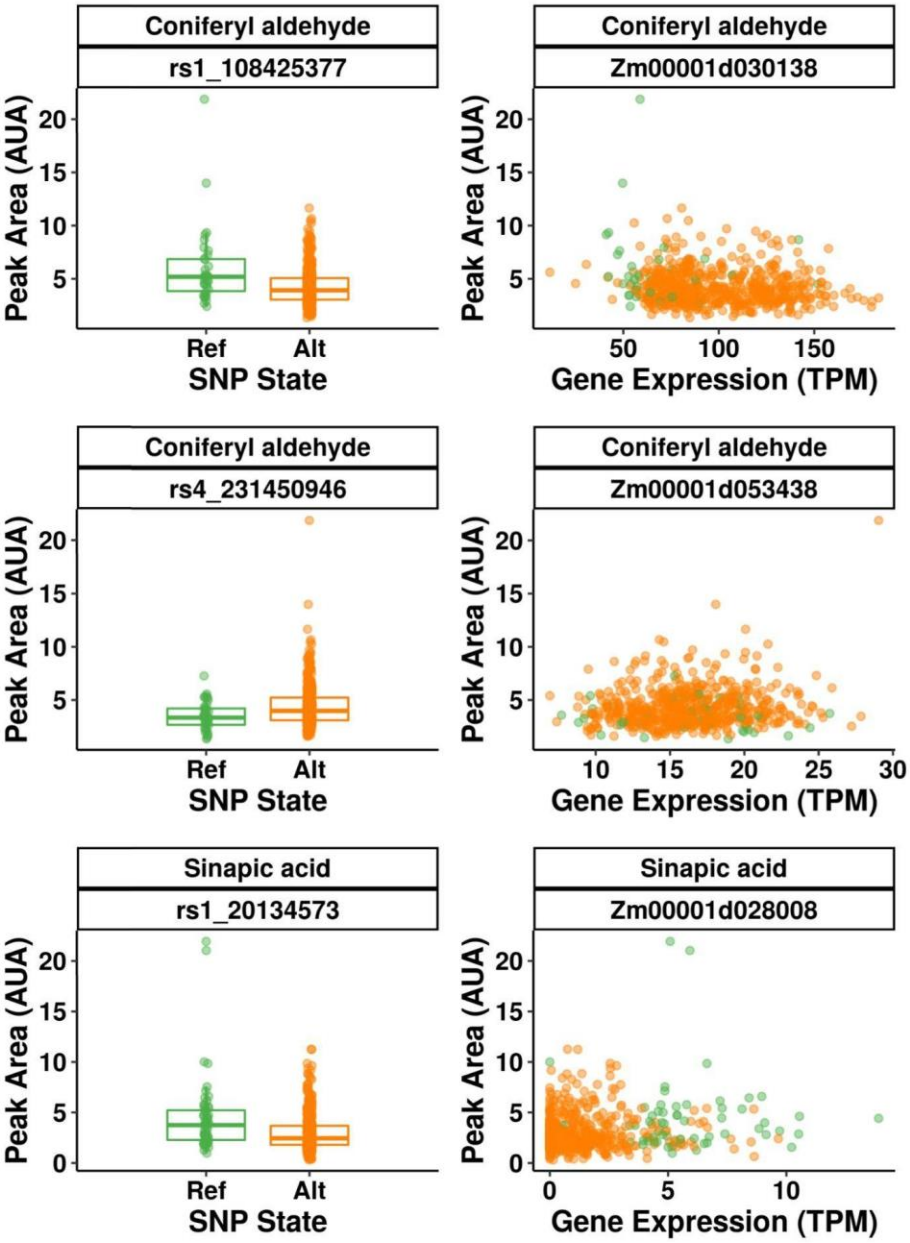


Figure 4.5: SNP states and gene expression profiles for three high confidence candidate genes identified for coniferyl aldehyde and sinapic acid. The SNPs shown in the left panels reside within the gene start and end sites of the genes shown in the corresponding right side panels. Inbred lines carrying the reference allele (B73v4) are plotted in green, and those carrying the alternate allele are shown in orange in all plots.

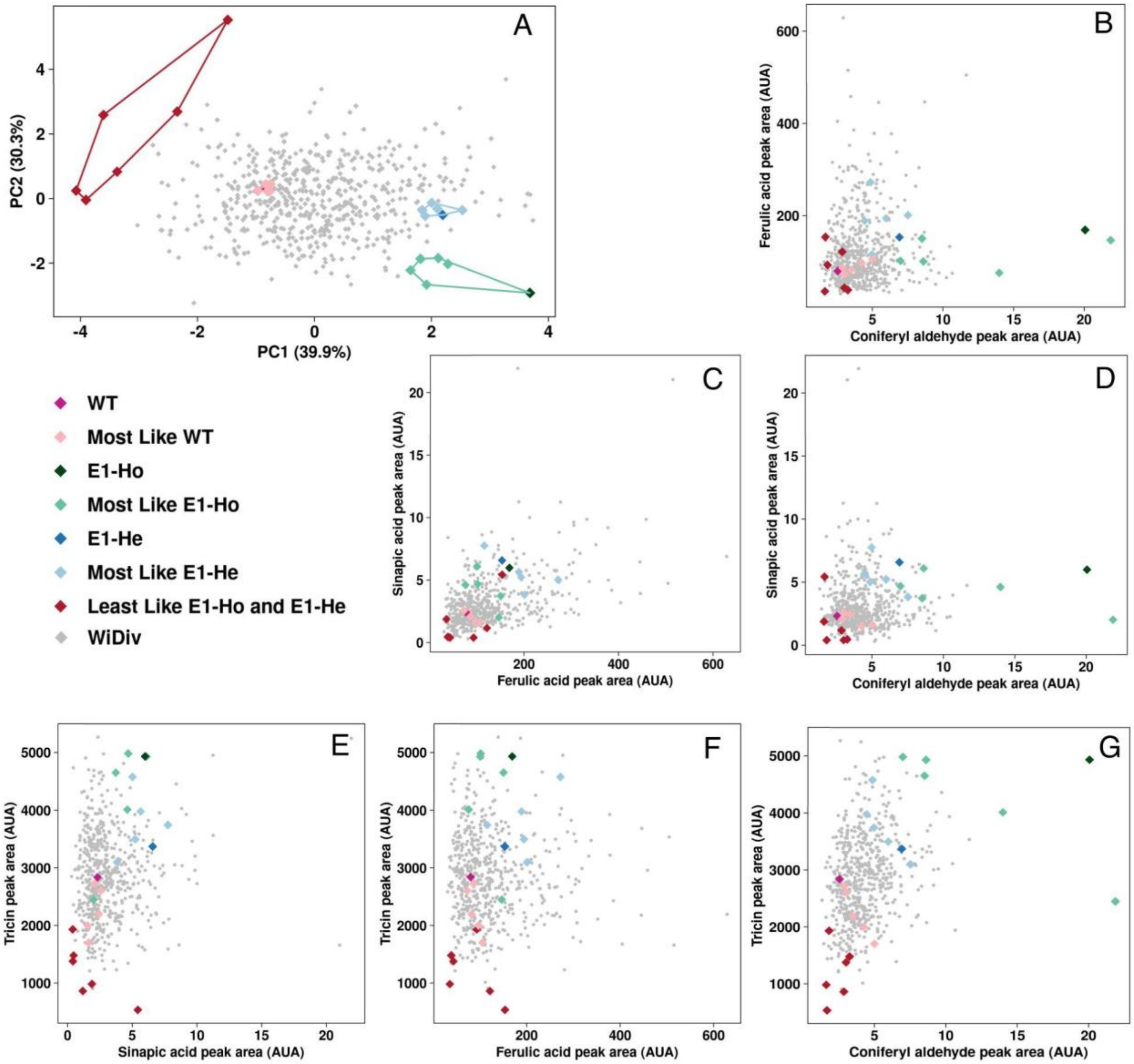
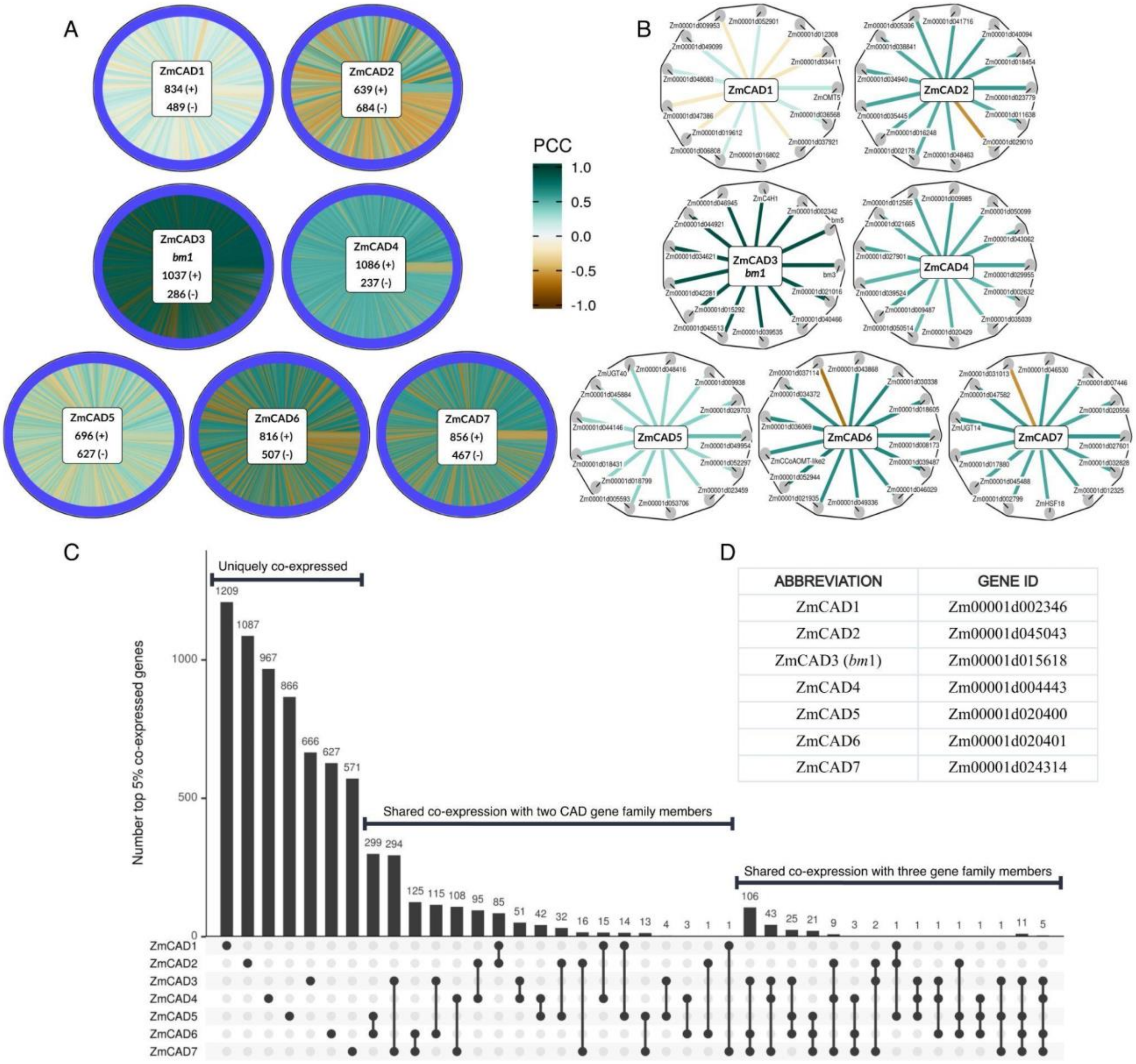


Figure 4.6: Phenolic metabolite profiles of bm1 RNAi lines and their WT counterpart overlaid on the range in natural variation for the same set of compounds that is captured by 597 WiDiv lines. (A) The first two PCs calculated from phenolic compound BLUPs, highlighting the WT and RNAi lines along with the top five WiDiv lines that are nearest (most like) or furthest from (least like) the WT and each RNAi line.



ABBREVIATION	GENE ID
ZmCAD1	Zm00001d002346
ZmCAD2	Zm00001d045043
ZmCAD3 (*bm*1)	Zm00001d015618
ZmCAD4	Zm00001d004443
ZmCAD5	Zm00001d020400
ZmCAD6	Zm00001d020401
ZmCAD7	Zm00001d024314

Figure 4.7: Genome-wide co-expression analysis of the ZmCAD gene family in maize. The Pearson correlation coefficient (PCC) between the expression of each gene family member and 26,095 protein-coding genes was calculated then PCCs sorted based on mutual rankings. (A) The top 5% of genes co-expressed with each CAD gene at the center. The number of genes with positive or negative PCC are denoted with a `+` or `-` sign respectively. (B) The top 0.05% of genes with each CAD gene at the center. (C) A summary of the number of genes co-expressed uniquely with each CAD gene, or shared among more than one gene family member. (D) A table connecting the B73v4 gene ID to each abbreviated gene symbol. The expression data used in this analysis was from RNA-seq of whole seedlings at the V1 developmental stage

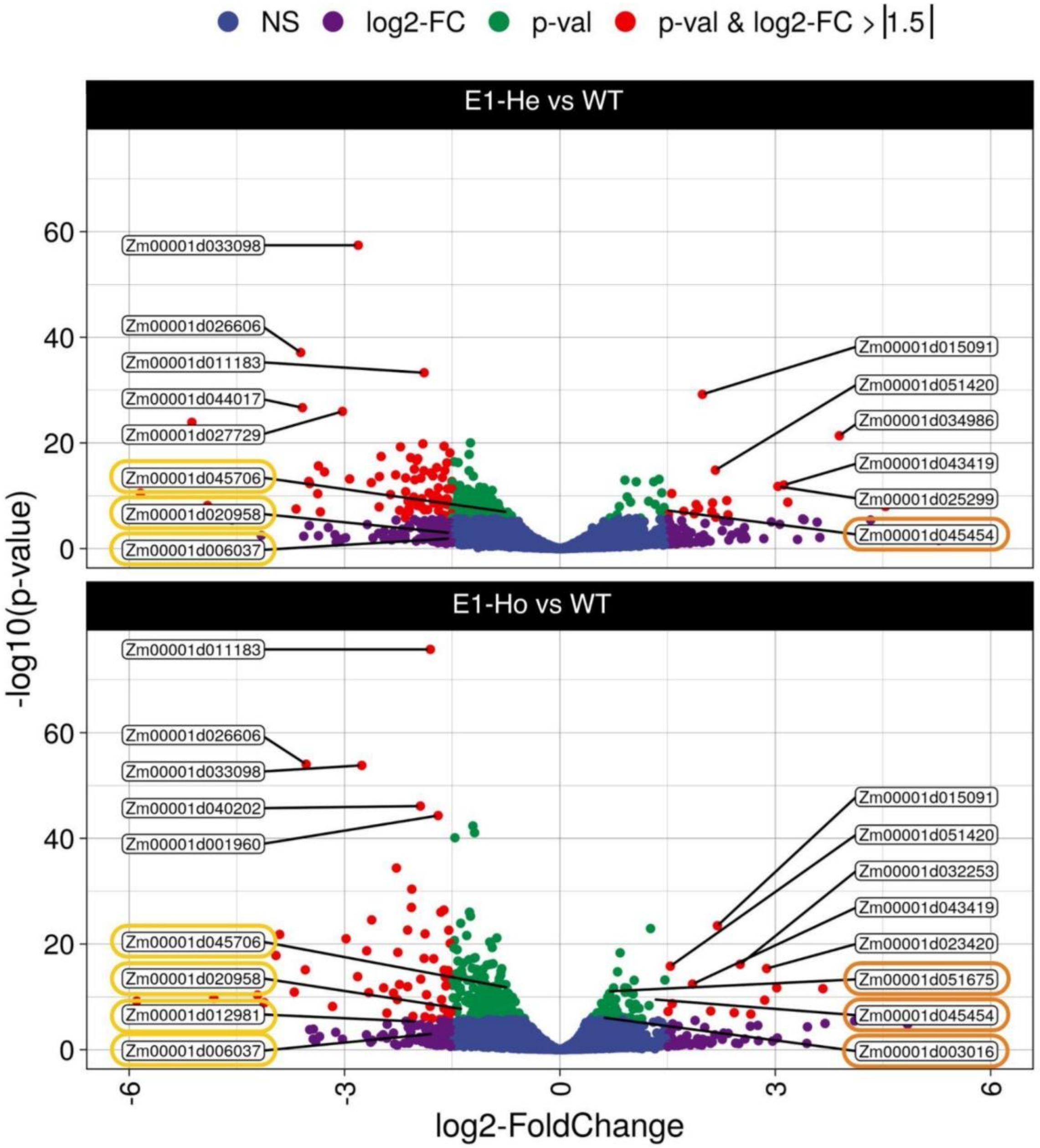


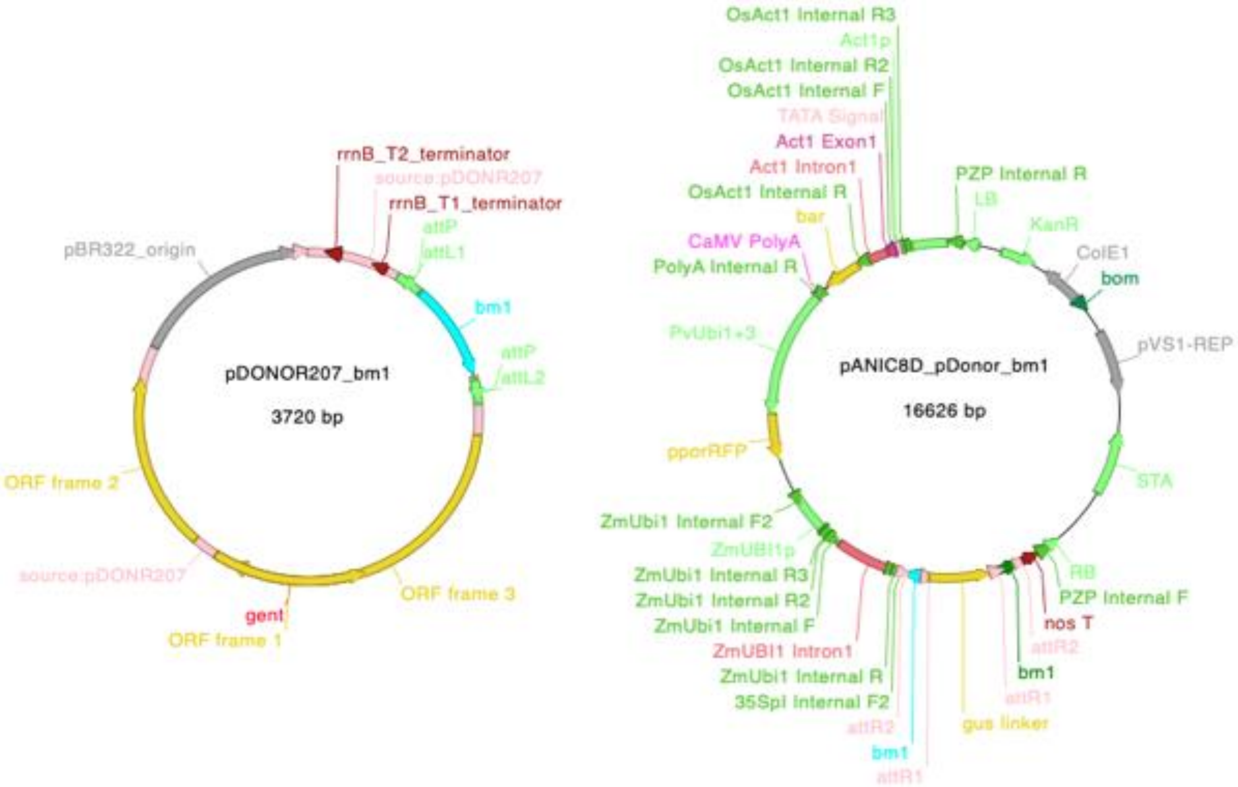
Figure 4.8: Volcano plots for differential gene expression analysis comparing gene expression changes associated with a disruption to bm1 by RNAi. The upper panel corresponds to the hemizygous RNAi line (E1-He), and the lower panel represents the homozygous RNAi line (E1-Ho). Labeled are the top 10 most significant gene expression changes (5 up regulated and 5 downregulated) between the RNAi line and each of the allelic states. Gene IDs outlined in yellow correspond to genes belonging to the HCALDH and CCR gene families, and genes outlined in orange are high-confidence candidate genes identified in both GWAS and TWAS as associated with at least one phenolic compound.

4.8 Supplemental Tables

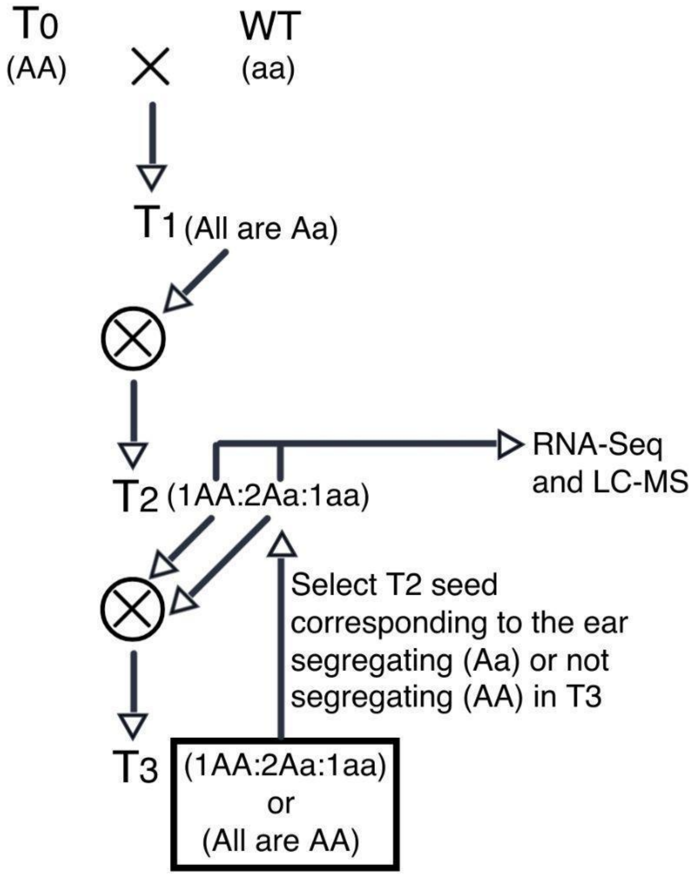
Supplemental Table 4.1 Summary of total RNA-Seq reads per sample (WT and RNAi lines), total reads, percent GC, and percent unique reads before trimming. Additionally, a summary of reads processed by Salmon, reads mapped to the B73 v4 transcriptome, and the percent of mapped reads.

Sample name	File name	Total reads	%GC	%unique reads	Reads processed	Reads mapped	%mapped
WT_rep1	S1_S58_forward.fastq.gz	42363163	46	27.42	42337686	24707068	58.36
WT_rep1	S1_S58_reverse.fastq.gz	42363163	46	29.34			
WT_rep2	S2_S65_forward.fastq.gz	39643008	45	26.50	39612697	23634486	59.66
WT_rep2	S2_S65_reverse.fastq.gz	39643008	45	28.87			
WT_rep3	S3_S66_forward.fastq.gz	40104076	46	27.47	40037211	23742565	59.30
WT_rep3	S3_S66_reverse.fastq.gz	40104076	46	29.78			
E1-He_rep1	S4_S67_forward.fastq.gz	34106264	45	34.26	34085194	21058702	61.78
E1-He_rep1	S4_S67_reverse.fastq.gz	34106264	46	36.29			
E1-He_rep2	S5_S68_forward.fastq.gz	40436258	46	32.31	40403297	24670341	61.06
E1-He_rep2	S5_S68_reverse.fastq.gz	40436258	46	34.65			
E1-He_rep3	S6_S69_forward.fastq.gz	39627961	46	31.13	39598375	23215278	58.63
E1-He_rep3	S6_S69_reverse.fastq.gz	39627961	46	33.07			
E1-Ho_rep1	S7_S70_forward.fastq.gz	42883540	45	31.20	42850760	26254401	61.27
E1-Ho_rep1	S7_S70_reverse.fastq.gz	42883540	46	33.46			
E1-Ho_rep2	S8_S71_forward.fastq.gz	41892806	47	25.91	41856850	21454064	51.26
E1-Ho_rep2	S8_S71_reverse.fastq.gz	41892806	48	27.62			
E1-Ho_rep3	S9_S72_forward.fastq.gz	44837017	46	29.87	44814886	26913077	60.05
E1-Ho_rep3	S9_S72_reverse.fastq.gz	44837017	46	32.00			
E2-He_rep1	S10_S59_forward.fastq.gz	50260482	45	27.09	50216310	31094926	61.92
E2-He_rep1	S10_S59_reverse.fastq.gz	50260482	45	29.55			
E2-He_rep2	S11_S60_forward.fastq.gz	40157443	45	28.73	40139241	24559681	61.19
E2-He_rep2	S11_S60_reverse.fastq.gz	40157443	45	31.30			
E2-He_rep3	S12_S61_forward.fastq.gz	42028148	45	31.10	42004872	26926587	64.10
E2-He_rep3	S12_S61_reverse.fastq.gz	42028148	45	33.90			
E2-Ho_rep1	S13_S62_forward.fastq.gz	39906535	45	31.31	39882574	25892180	64.92
E2-Ho_rep1	S13_S62_reverse.fastq.gz	39906535	45	34.03			
E2-Ho_rep2	S14_S63_forward.fastq.gz	39370611	46	32.02	39339554	24916941	63.34
E2-Ho_rep2	S14_S63_reverse.fastq.gz	39370611	46	34.61			
E2-Ho_rep3	S15_S64_forward.fastq.gz	37255467	45	28.04	37233083	22704561	60.98
E2-Ho_rep3	S15_S64_reverse.fastq.gz	37255467	45	30.41			

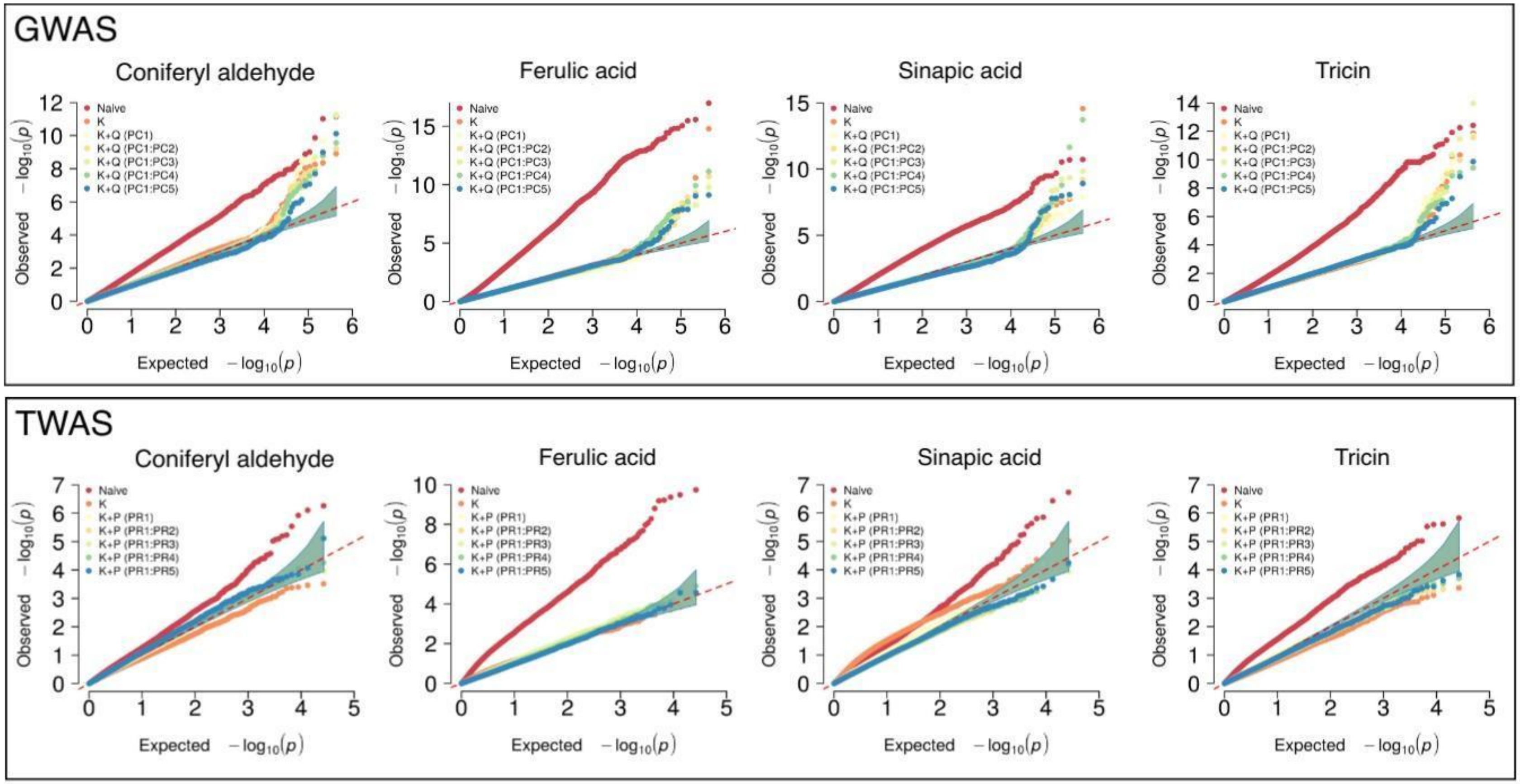
4.9 Supplemental Figures



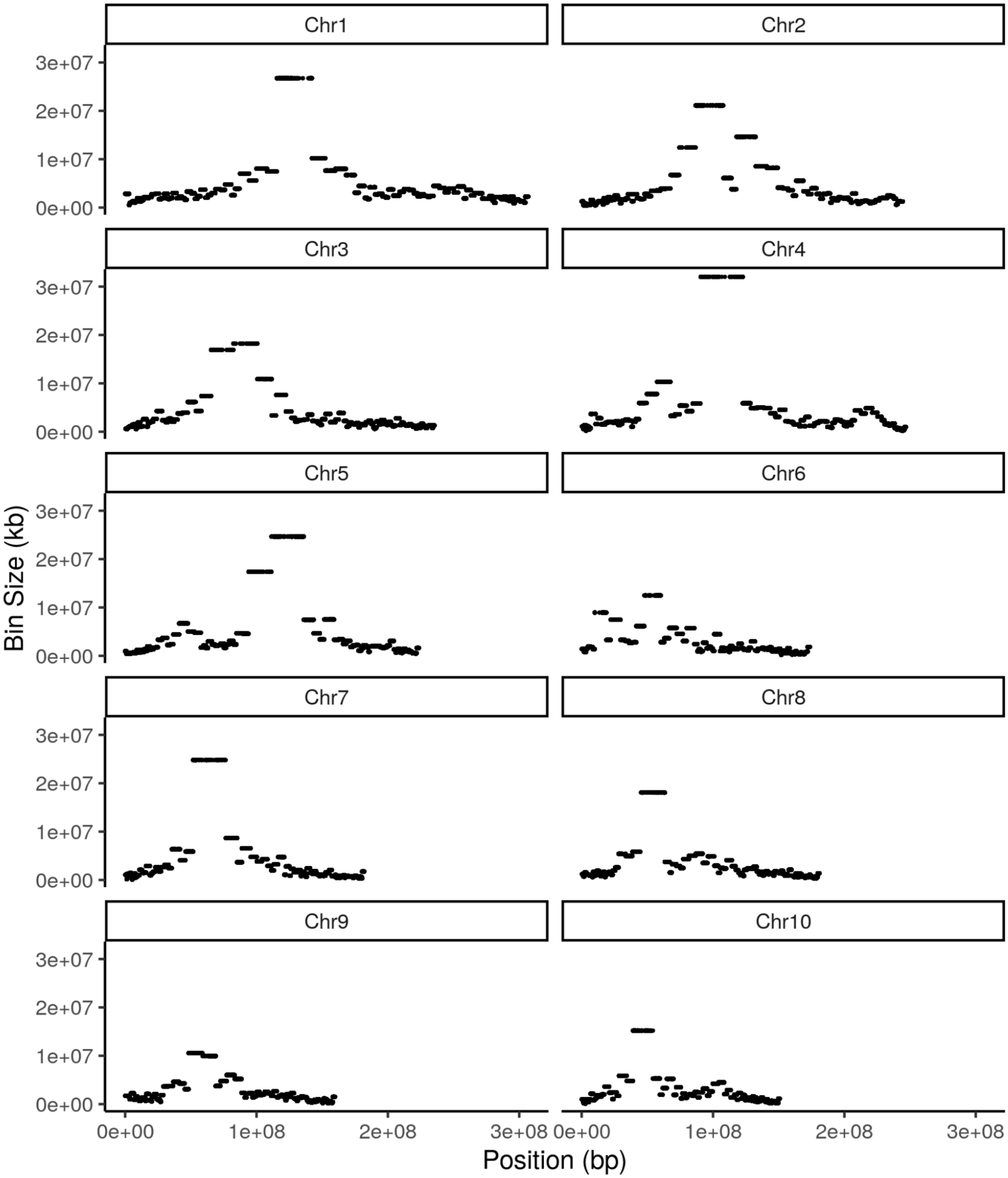
Supplemental Figure 4.1: Entry vector (pDONOR207) and expression (pANIC8D) construct used for Agrobacterium-mediated transformation to produce bm1 RNAi lines.



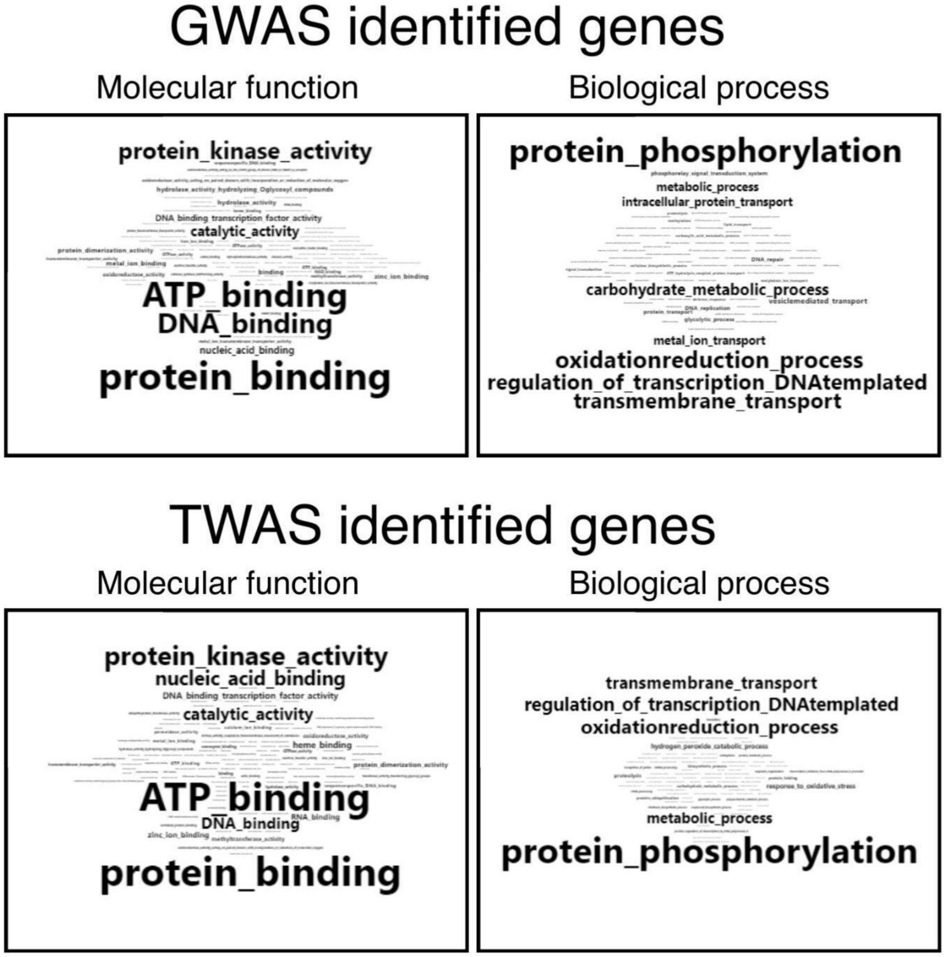
Supplemental Figure 4.2: Crossing scheme used for producing hemizygous (Aa) and homozygous (AA) bm1 RNAi lines. For each generation, the expected segregation ratio of the bm1 RNAi transgene on individual ears is shown in parenthesis.



Supplemental Figure 4.3: Quantile-Quantile plots for GWAS and TWAS of each phenolic compound. For GWAS, the models tested included a naive model, then kinship (K) and sequential increase in the number of SNP calculated principal components (PCs) included in the model. For TWAS, the models tested were the same as GWAS, but with a sequential increase in gene expression calculated peer factors (PRs) included in the model.



Supplemental Figure 4.4: Chromosomal bins generated using 428,287 SNPs. The chromosome bin size is a function of marker density since each chromosome is split into 100 bins based on quantiles of the physical positions of SNPs.



Supplemental Figure 4.5: GO term word clouds for genes identified either in GWAS or TWAS. Larger size indicates increased frequency of the term.

4.10 Supplemental Files

Supplemental File 4.1: Decoder file connecting WiDiv genotype names to LC-MS phenolic compound data (Deposited as a separate file with this thesis).

Supplemental File 4.2: DEG results from DESeq2 for bm1 RNAi lines (Deposited as a separate file with this thesis).

4.11 References

Alam, M. A., Subhan, N., Hossain, H., Hossain, M., Reza, H. M., Rahman, M. M., & Ullah, M. O. (2016). Hydroxycinnamic acid derivatives: A potential class of natural compounds for the management of lipid metabolism and obesity. *Nutrition & Metabolism*, *13*(1), 27. https://doi.org/10.1186/s12986-016-0080-3

Azodi, C. B., Pardo, J., VanBuren, R., de los Campos, G., & Shiu, S.-H. (2020). Transcriptome-Based Prediction of Complex Traits in Maize. *The Plant Cell*, *32*(1), 139–151. https://doi.org/10.1105/tpc.19.00332

Barrière, Y., Argillier, O., Barrière, Y., & Brown-midrib, O. A. (1993). Brown-midrib genes of maize: A review. *Agronomie*, *13*(10), 865–876. https://hal.archives-ouvertes.fr/hal-00885517

Barrière, Y., Courtial, A., Chateigner-Boutin, A. L., Denoue, D., & Grima-Pettenati, J. (2015). Breeding maize for silage and biofuel production, an illustration of a step forward with the genome sequence. *Plant Science*, *242*, 310–329. https://doi.org/10.1016/j.plantsci.2015.08.007

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

Boerjan, W., Ralph, J., & Baucher, M. (2003). Lignin Biosynthesis. *Annual Review of Plant Biology*, *54*(1), 519–546. https://doi.org/10.1146/annurev.arplant.54.031902.134938

Broad Institute. (2022). *Analysis pipelines for the GTEx Consortium and TOPMed* [Python]. Broad Institute. https://github.com/broadinstitute/gtex-pipeline/blob/f5b2abe0ad6a8618e19e89e72f3c8e0594c75308/qtl/README.md

del Río, J. C., Rencoret, J., Gutiérrez, A., Elder, T., Kim, H., & Ralph, J. (2020). Lignin Monomers from beyond the Canonical Monolignol Biosynthetic Pathway: Another Brick in the Wall. *ACS Sustainable Chemistry & Engineering*, *8*(13), 4997–5012. https://doi.org/10.1021/acssuschemeng.0c01109

Deng, Y., & Lu, S. (2017). Biosynthesis and Regulation of Phenylpropanoids in Plants. *Critical Reviews in Plant Sciences*, *36*(4), 257–290. https://doi.org/10.1080/07352689.2017.1402852

Domínguez-Robles, J., Martin, N. K., Fong, M. L., Stewart, S. A., Irwin, N. J., Rial-Hermida, M. I., Donnelly, R. F., & Larrañeta, E. (2019). Antioxidant PLA Composites Containing Lignin for 3D Printing Applications: A Potential Material for Healthcare Applications. *Pharmaceutics*, *11*(4), 165. https://doi.org/10.3390/pharmaceutics11040165

Fisher, R. A. (1938). *Statistical methods for research workers*. Edinburgh, Oliver and Boyd. http://archive.org/details/statisticalmethoe7fish

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology*, *54*(1), 357–374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Gage, J. L., White, M. R., Edwards, J. W., Kaeppler, S., & de Leon, N. (2018). Selection signatures underlying dramatic male inflorescence transformation during modern hybrid maize breeding. *Genetics*, *210*(3), 1125–1138. https://doi.org/10.1534/genetics.118.301487

Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., Nemhauser, J. L., Schmitz, R. J., & Gallavotti, A. (2018). The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nature Communications*, *9*(1), 4526. https://doi.org/10.1038/s41467-018-06977-6

Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z., & Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theoretical and Applied Genetics*, *129*(12), 2413–2427. https://doi.org/10.1007/s00122-016-2780-5

Hansey, C. N., Lorenz, A. J., & de Leon, N. (2010). Cell Wall Composition and Ruminant Digestibility ofVarious Maize Tissues Across Development. *Bioenergy Research*, *3*(3), 295–304. https://doi.org/10.1007/s12155-010-9100-8

Heazlewood, J. L., Howell, K. A., & Millar, A. H. (2003). Mitochondrial complex I from Arabidopsis and rice: Orthologs of mammalian and fungal components coupled with plant-specific subunits. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1604*(3), 159–169. https://doi.org/10.1016/S0005-2728(03)00045-8

Hirsch, C. N., Foerster, J. M., Johnson, J. M., Sekhon, R. S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M. A., Barry, K., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2014). Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell*, *26*(1), 121–135. https://doi.org/10.1105/tpc.113.119982

Hoopes, G. M., Hamilton, J. P., Wood, J. C., Esteban, E., Pasha, A., Vaillancourt, B., Provart, N. J., & Buell, C. R. (2018). *Data from: An updated gene atlas for maize reveals organ-specific and stress-induced genes* (Version 1, p. 377983485 bytes) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.5P58Q34

Hoopes, G. M., Hamilton, J. P., Wood, J. C., Esteban, E., Pasha, A., Vaillancourt, B., Provart, N. J., & Buell, C. R. (2019). An updated gene atlas for maize reveals organ-specific and stress-induced genes. *The Plant Journal*, *97*(6), 1154–1167. https://doi.org/10.1111/tpj.14184

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K. L., Wolfgruber, T. K., May, M. R., Springer, N. M., … Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*(7659), 524–527. https://doi.org/10.1038/nature22971

Jung, H. G., Morrison, T. A., & Buxton, D. R. (1998). Degradability of cell-wall polysaccharides in maize internodes during stalk development. *Crop Science*, *38*(4), 1047–1051. https://doi.org/10.2135/cropsci1998.0011183X003800040027x

Karlen, S. D., Fasahati, P., Mazaheri, M., Serate, J., Smith, R. A., Sirobhushanam, S., Chen, M., Tymokhin, V. I., Cass, C. L., Liu, S., Padmakshan, D., Xie, D., Zhang, Y., McGee, M. A., Russell, J. D., Coon, J. J., Kaeppler, H. F., de Leon, N., Maravelias, C. T., … Ralph, J. (2020). Assessing the Viability of Recovery of Hydroxycinnamic Acids from Lignocellulosic Biorefinery Alkaline Pretreatment Waste Streams. *ChemSusChem*, 1–14. https://doi.org/10.1002/cssc.201903345

Karlen, S. D., Zhang, C., Peck, M. L., Smith, R. A., Padmakshan, D., Helmich, K. E., Free, H. C. A., Lee, S., Smith, B. G., Lu, F., Sedbrook, J. C., Sibout, R., Grabber, J. H., Runge, T. M., Mysore, K. S., Harris, P. J., Bartley, L. E., & Ralph, J. (2016). Monolignol ferulate conjugates are naturally incorporated into plant lignins. *Science Advances*, *2*(10), e1600393. https://doi.org/10.1126/sciadv.1600393

Kassambara, A. (2020, June 27). *ggpubr: "ggplot2" Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr

Končitíková, R., Vigouroux, A., Kopečná, M., Andree, T., Bartoš, J., Šebela, M., Moréra, S., & Kopečný, D. (2015). Role and structural characterization of plant aldehyde dehydrogenases from family 2 and family 7. *Biochemical Journal*, *468*(1), 109–123. https://doi.org/10.1042/BJ20150009

Kremling, K. A. G., Diepenbrock, C. H., Gore, M. A., Buckler, E. S., & Bandillo, N. B. (2019). Transcriptome-wide association supplements genome-wide association in Zea mays. *G3: Genes, Genomes, Genetics*, *9*(9), 3023–3033. https://doi.org/10.1534/g3.119.400549

Kumar, A., Kumar, V., Kumar, A., Antonio, F., Antunes, F., & Silvério, S. (2018). Bioresource Technology The path forward for lignocellulose biorefineries: Bottlenecks, solutions, and perspective on commercialization. *Bioresource Technology*, *264*(June), 370–381. https://doi.org/10.1016/j.biortech.2018.06.004

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Laftah, W. A., Hashim, S., & Ibrahim, A. N. (2011). Polymer Hydrogels: A Review. *Polymer-Plastics Technology and Engineering*, *50*(14), 1475–1486. https://doi.org/10.1080/03602559.2011.593082

Lan, W., Morreel, K., Lu, F., Rencoret, J., Carlos del Río, J., Voorend, W., Vermerris, W., Boerjan, W., & Ralph, J. (2016). Maize Tricin-Oligolignol Metabolites and Their Implications for Monocot Lignification. *Plant Physiology*, *171*(2), 810–820. https://doi.org/10.1104/pp.16.02012

Lee, S.-W., Seo, J. M., Lee, M.-K., Chun, J.-H., Antonisamy, P., Arasu, M. V., Suzuki, T., Al-Dhabi, N. A., & Kim, S.-J. (2014). Influence of different LED lamps on the production of phenolic compounds in common and Tartary buckwheat sprouts. *Industrial Crops and Products*, *54*, 320–326. https://doi.org/10.1016/j.indcrop.2014.01.024

Lei, L. (2017). Lignin evolution: Invasion of land. *Nature Plants*, *3*(4), 1–1. https://doi.org/10.1038/nplants.2017.42

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. *BMC Genetics*, *13*(1), 100–100. https://doi.org/10.1186/1471-2156-13-100

Liu, X., Van Acker, R., Voorend, W., Pallidis, A., Goeminne, G., Pollier, J., Morreel, K., Kim, H., Muylle, H., Bosio, M., Ralph, J., Vanholme, R., & Boerjan, W. (2021). Rewired phenolic metabolism and improved saccharification efficiency of a Zea mays cinnamyl alcohol dehydrogenase 2 (zmcad2) mutant. *The Plant Journal: For Cell and Molecular Biology*, *105*(5), 1240–1257. https://doi.org/10.1111/tpj.15108

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

Lowe, B. A., Way, Æ. M. M., Kumpf, J. M., Rout, Æ. J., Warner, Æ. D., Johnson, R., Armstrong, Æ. C. L., Spencer, M. T., & Chomet, Æ. P. S. (2006). *Marker assisted breeding for transformability in maize*. 229–239. https://doi.org/10.1007/s11032-006-9031-4

Mann, D. G. J., LaFayette, P. R., Abercrombie, L. L., King, Z. R., Mazarei, M., Halter, M. C., Poovaiah, C. R., Baxter, H., Shen, H., Dixon, R. A., Parrott, W. A., & Neal Stewart Jr, C. (2012). Gateway-compatible vectors for high-throughput gene functional analysis in switchgrass (Panicum virgatum L.) and other monocot species. *Plant Biotechnology Journal*, *10*(2), 226–236. https://doi.org/10.1111/j.1467-7652.2011.00658.x

Mansell, R. L., Babbel, G. R., & Zenk, M. H. (1976). Multiple forms and specificity of coniferyl alcohol dehydrogenase from cambial regions of higher plants. *Phytochemistry*, *15*(12), 1849–1853. https://doi.org/10.1016/S0031-9422(00)88829-9

Mao, C., Zhu, Y., Cheng, H., Yan, H., Zhao, L., Tang, J., Ma, X., & Mao, P. (2018). Nitric Oxide Regulates Seedling Growth and Mitochondrial Responses in Aged Oat Seeds. *International Journal of Molecular Sciences*, *19*(4), 1052. https://doi.org/10.3390/ijms19041052

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

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., Robin Buell, C., de Leon, N., & Kaeppler, S. M. (2019). Genome-wide association analysis of stalk biomass and anatomical traits in maize. *BMC Plant Biology*, *19*(1), 1–17. https://doi.org/10.1186/s12870-019-1653-x

Mittasch, J., Böttcher, C., Frolov, A., Strack, D., & Milkowski, C. (2013). Reprogramming the Phenylpropanoid Metabolism in Seeds of Oilseed Rape by Suppressing the Orthologs of REDUCED EPIDERMAL FLUORESCENCE1. *Plant Physiology*, *161*(4), 1656–1669. https://doi.org/10.1104/pp.113.215491

Morrison, T. A., Kessler, J. R., Hatfield, R. D., & Buxton, D. R. (1994). Activity of two lignin biosynthesis enzymes during development of a maize internode. *Journal of the Science of Food and Agriculture*, *65*(2), 133–139. https://doi.org/10.1002/jsfa.2740650202

Nair, R. B., Bastress, K. L., Ruegger, M. O., Denault, J. W., & Chapple, C. (2004). The Arabidopsis thaliana REDUCED EPIDERMAL FLUORESCENCE1 Gene Encodes an Aldehyde Dehydrogenase Involved in Ferulic Acid and Sinapic Acid Biosynthesis. *The Plant Cell*, *16*(2), 544–554. https://doi.org/10.1105/tpc.017509

Obayashi, T., & Kinoshita, K. (2009). Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. *DNA Research*, *16*(5), 249–260. https://doi.org/10.1093/dnares/dsp016

Oliver, S. (2000). Guilt-by-association goes global. *Nature*, *403*(6770), 601–602. https://doi.org/10.1038/35001165

Othibeng, K., Nephali, L., Ramabulana, A.-T., Steenkamp, P., Petras, D., Kang, K. B., Opperman, H., Huyser, J., & Tugizimana, F. (2021). A Metabolic Choreography of Maize Plants Treated with a Humic Substance-Based Biostimulant under Normal and Starved Conditions. *Metabolites*, *11*(6), 403. https://doi.org/10.3390/metabo11060403

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Prive, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics*, *34*(16), 2781–2787. https://doi.org/10.1093/bioinformatics/bty185

R Core Team. (2020). R Core Team R: A language and environment for statistical computing. *Foundation for Statistical Computing*.

Ralph, J., Lundquist, K., Brunow, G., Lu, F., Kim, H., Schatz, P. F., Marita, J. M., Hatfield, R. D., Ralph, S. A., Christensen, J. H., & Boerjan, W. (2004). Lignins: Natural polymers from oxidative coupling of 4-hydroxyphenyl- propanoids. *Phytochemistry Reviews*, *3*(1–2), 29–60. https://doi.org/10.1023/B:PHYT.0000047809.65444.a4

Ricci, W. A., Lu, Z., Ji, L., Marand, A. P., Ethridge, C. L., Murphy, N. G., Noshay, J. M., Galli, M., Mejía-Guerra, M. K., Colomé-Tatché, M., Johannes, F., Rowley, M. J., Corces, V. G., Zhai, J., Scanlon, M. J., Buckler, E. S., Gallavotti, A., Springer, N. M., Schmitz, R. J.,

& Zhang, X. (2019). Widespread Long-range Cis-Regulatory Elements in the Maize Genome. *Nature Plants*, *5*(12), 1237. https://doi.org/10.1038/s41477-019-0547-0

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

Schrag, T. A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., & Melchinger, A. E. (2018). Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. *Genetics*, *208*(4), 1373–1385. https://doi.org/10.1534/genetics.117.300374

Sekhon, R. S., Lin, H., Childs, K. L., Hansey, C. N., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2011). Genome-wide atlas of transcription during maize development. *The Plant Journal: For Cell and Molecular Biology*, *66*(4), 553–563. https://doi.org/10.1111/j.1365-313X.2011.04527.x

Seo, J.-M., Arasu, M. V., Kim, Y.-B., Park, S. U., & Kim, S.-J. (2015). Phenylalanine and LED lights enhance phenolic compound production in Tartary buckwheat sprouts. *Food Chemistry*, *177*, 204–213. https://doi.org/10.1016/j.foodchem.2014.12.094

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591–611. https://doi.org/10.2307/2333709

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507. https://doi.org/10.1038/nprot.2011.457

Stucker, D. S., & Hallauer, A. R. (1992). Genetic Variability as Affected by Selection in lowa Stiff Stalk Synthetic Maize. *Journal of Heredity*, *83*(6), 410–418. https://doi.org/10.1093/oxfordjournals.jhered.a111243

Tu, X., Mejía-Guerra, M. K., Valdes Franco, J. A., Tzeng, D., Chu, P. Y., Shen, W., Wei, Y., Dai, X., Li, P., Buckler, E. S., & Zhong, S. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nature Communications*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-18832-8

Vanholme, R., Morreel, K., Darrah, C., Oyarce, P., Grabber, J. H., Ralph, J., & Boerjan, W. (2012). Metabolic engineering of novel lignin in biomass crops. *New Phytologist*, *196*(4), 978–1000. https://doi.org/10.1111/j.1469-8137.2012.04337.x

Vignols, F. (1995). The brown midrib3 (bm3) Mutation in Maize Occurs in the Gene Encoding Caffeic Acid O-Methyltransferase. *The Plant Cell Online*, *7*(4), 407–416. https://doi.org/10.1105/tpc.7.4.407

Vogt, T. (2010). Phenylpropanoid Biosynthesis. *Molecular Plant*, *3*(1), 2–20. https://doi.org/10.1093/mp/ssp106

Wallace, J. G., Bradbury, P. J., Zhang, N., Gibon, Y., Stitt, M., & Buckler, E. S. (2014). Association Mapping across Numerous Traits Reveals Patterns of Functional Variation in Maize. *PLOS Genetics*, *10*(12), e1004845. https://doi.org/10.1371/journal.pgen.1004845

Wang, P., Dudareva, N., Morgan, J. A., & Chapple, C. (2015). Genetic manipulation of lignocellulosic biomass for bioenergy. *Current Opinion in Chemical Biology*, *29*, 32–39. https://doi.org/10.1016/j.cbpa.2015.08.006

Westhues, M., Schrag, T. A., Heuer, C., Thaller, G., Utz, H. F., Schipprack, W., Thiemann, A., Seifert, F., Ehret, A., Schlereth, A., Stitt, M., Nikoloski, Z., Willmitzer, L., Schön, C. C., Scholten, S., & Melchinger, A. E. (2017). Omics-based hybrid prediction in maize. *Theoretical and Applied Genetics*, *130*(9), 1927–1939. https://doi.org/10.1007/s00122-017-2934-0

White, M. R., Mikel, M. A., de 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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williams, E. J. B., & Bowles, D. J. (2004). Coexpression of Neighboring Genes in the Genome of Arabidopsis thaliana. *Genome Research*, *14*(6), 1060–1067. https://doi.org/10.1101/gr.2131104

Xiong, W., Li, Y., Wu, Z., Ma, L., Liu, Y., Qin, L., Liu, J., Hu, Z., Guo, S., Sun, J., Yang, G., Chai, M., Zhang, C., Lu, X., & Fu, C. (2020). Characterization of Two New brown midrib1 Mutations From an EMS-Mutagenic Maize Population for Lignocellulosic Biomass Utilization. *Frontiers in Plant Science*, *11*. https://www.frontiersin.org/article/10.3389/fpls.2020.594798

Yan, X., Liu, J., Kim, H., Liu, B., Huang, X., Yang, Z., Lin, Y.-C. J., Chen, H., Yang, C., Wang, J. P., Muddiman, D. C., Ralph, J., Sederoff, R. R., Li, Q., & Chiang, V. L. (2019). CAD1 and CCR2 protein complex formation in monolignol biosynthesis in Populus trichocarpa. *New Phytologist*, *222*(1), 244–260. https://doi.org/10.1111/nph.15505

Yang, F., Li, W., Jiang, N., Yu, H., Morohashi, K., Ouma, W. Z., Morales-Mantilla, D. E., Gomez-Cano, F. A., Mukundi, E., Prada-Salcedo, L. D., Velazquez, R. A., Valentin, J., Mejía-Guerra, M. K., Gray, J., Doseff, A. I., & Grotewold, E. (2017). A Maize Gene Regulatory Network for Phenolic Metabolism. *Molecular Plant*, *10*(3), 498–515. https://doi.org/10.1016/j.molp.2016.10.020

Yang, Y., Saand, M. A., Huang, L., Abdelaal, W. B., Zhang, J., Wu, Y., Li, J., Sirohi, M. H., & Wang, F. (2021). Applications of Multi-Omics Technologies for Crop Improvement. *Frontiers in Plant Science*, *12*, 1846. https://doi.org/10.3389/fpls.2021.563953

Yoon, S., Baik, B., Park, T., & Nam, D. (2021). Powerful p-value combination methods to detect incomplete association. *Scientific Reports*, *11*(1), 6980. https://doi.org/10.1038/s41598-021-86465-y

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). 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

Zhang, X., & Jonassen, I. (2020). RASflow: An RNA-Seq analysis workflow with Snakemake. *BMC Bioinformatics*, *21*(1), 110. https://doi.org/10.1186/s12859-020-3433-x

Zhou, P., & Springer, N. M. (2020). *Data for: Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions* [Data set]. https://doi.org/10.13020/p3g0-3170

Zhou, S., Kremling, K. A., Bandillo, N., Richter, A., Zhang, Y. K., Ahern, K. R., Artyukhin, A. B., Hui, J. X., Younkin, G. C., Schroeder, F. C., Buckler, E. S., & Jander, G. (2019). Metabolome-Scale Genome-Wide Association Studies Reveal Chemical Diversity and Genetic Control of Maize Specialized Metabolites. *The Plant Cell*, *31*(5), 937–955. https://doi.org/10.1105/tpc.18.00772